Time Series Analysis

Lecture Review

Dr. Kalyan N
Assistant Professor
Dept. of CSE (Data Science)
B.M.S College of Engineering
Bengaluru - 560019.
kalyan.cds@bmsce.ac.in
Homepage

October, 2024.

Contents

Module - 1

Chapter - 1

A time series is a sequence of statistical data organized according to the time of occurrence or in chronological order. The numerical data collected at various points in time, forming a set of observations, is referred to as a time series. In time series analysis, current data within a series can be compared with past data from the same series. Additionally, the progression of two or more series over time can be compared. These comparisons can provide valuable insights for individual businesses. Time series analysis is crucial in fields such as economics, statistics, and commerce.

1 Time Series Data

A time series consists of observations made at specific time intervals and arranged in chronological order. For example, tracking agricultural production, sales, or National Income over a span of 3 to 5 years constitutes a time series. It is essentially a sequence of quantitative readings recorded at regular intervals, which could be hourly, daily, weekly, monthly, or annually. Examples of time series include hourly temperature readings, daily shop sales, weekly market sales, monthly production figures, yearly agricultural outputs, and population growth over ten years. Analyzing a time series involves comparing past data with current data to forecast future trends and evaluate past performance. The focus of time series analysis is on understanding chronological variations. Key requirements for a time series are:

- The time intervals between observations should be as consistent as possible.
- The dataset must be homogeneous.
- Data should be collected over an extended period.

Symbolically, if t represents time and y_t denotes the value at time t, then the paired values (t, y_t) constitute the time series data. Ex 1: Production of rice in Karnataka for the period from 2010-11 to 2016-17.

Table 1: Production of rice in Karnataka (in ‘000 metric tons)

Year	**Production**
2010-11	800
2011-12	950
2012-13	870
2013-14	920
2014-15	860
2016-17	720

1.1 Purpose

Time series analysis is crucial for understanding historical data and forecasting future trends, which aids managers and policymakers in making informed decisions. By quantifying key features and random variations in data, time series methods have become widely applicable across government, industry, and commerce, especially with advances in computing power. The Kyoto Protocol, an amendment to the United Nations Framework Convention on Climate Change, was signed in December 1997 and came into effect on February 16, 2005. The rationale for reducing greenhouse gas emissions involves a blend of scientific data, economic considerations, and time series analysis. The decisions made in the coming years will have significant implications for the planet’s future.

In 2006, Singapore Airlines expanded its fleet by ordering twenty Boeing 787-9s and expressing intent to purchase twenty-nine Airbus planes, including twenty A350s and nine A380s (superjumbos). This expansion was guided by time series analysis of passenger trends and strategic corporate planning to maintain or enhance market share. Time

series methods are also employed in everyday operational decisions. For instance, UK gas suppliers must place orders for offshore gas one day in advance. The variation from the seasonal average is influenced by temperature and, to a lesser extent, wind speed. Time series analysis helps forecast demand by adjusting the seasonal average with one-day-ahead weather forecasts.
Additionally, time series models underpin many computer simulations. Examples include evaluating inventory control strategies using simulated demand series, comparing wave power device designs with simulated sea states, and simulating daily rainfall to assess the long-term environmental impacts of proposed water management policies.

1.2 Time series

In many fields, including science, engineering, and commerce, variables are measured sequentially over time. For instance, reserve banks track daily interest and exchange rates, governments report annual GDP figures, and meteorological offices log rainfall at various locations. When data are collected at regular intervals, they form a time series. A historical time series is created from observations recorded at fixed intervals. In this context, time series are often treated as realizations of sequences of random variables, known as discrete-time stochastic processes or time series models. Our focus will be on applying these models using R to fit data and perform analysis.

Time series data often exhibit trends and seasonal variations that can be modeled mathematically. Additionally, observations close in time are typically correlated. Time series analysis aims to explain this correlation and other data features using statistical models. Once a model is fitted, it can be used to forecast future values, conduct statistical tests, and summarize the main characteristics of the data, aiding decision-making.Sampling intervals impact data quality. Aggregated data, like daily tourist arrivals, or sampled data, such as daily stock prices, need appropriate intervals to accurately reflect the original signal. In high-frequency trading or signal processing, continuous signals are sampled at very high rates to create time series for detailed analysis.

1.2.1 Uses of Time series

The analysis of time series is of great significance not only to economists and business people but also to scientists, astronomers, geologists, sociologists, biologists, and researchers. This is due to the following reasons:

- It helps in understanding past behavior.
- It assists in planning future operations.
- It aids in evaluating current accomplishments.
- It facilitates comparison.

1.2.2 Plots

Visualizing time series data is crucial for identifying patterns and trends. Common types of plots include:

- **Line Plot:** Displays data points connected by lines to show changes over time. Useful for identifying trends and seasonal patterns.
- **Scatter Plot:** Plots individual data points to observe the relationship between two variables or to identify patterns and outliers.
- **Bar Plot:** Represents data with bars, helpful for comparing discrete time periods or categories.
- **Histogram:** Shows the distribution of data over specified intervals, useful for understanding the frequency of values.
- **Box Plot:** Displays the distribution of data based on quartiles, highlighting median, and potential outliers.

1.2.3 Trends

Trends refer to the long-term movement or direction in the data over a period. Identifying trends helps in understanding the overall pattern:

- **Upward Trend:** Indicates a general increase in values over time.
- **Downward Trend:** Shows a general decrease in values.
- **Stationary Trend:** The data fluctuates around a constant mean without a long-term trend.

1.2.4 Seasonal Variation

Variations refer to the deviations from the trend and can be categorized into:

- **Seasonal Variations:** Regular patterns that repeat at consistent intervals, such as monthly or quarterly.
- **Cyclical Variations:** Fluctuations that occur over longer periods, influenced by economic or business cycles.
- **Irregular Variations:** Unpredictable changes due to unforeseen events or anomalies that do not follow a pattern.

Understanding these components allows for effective analysis and forecasting of time series data.

1.3 Decomposition of Series

1.3.1 Notation

The analysis so far has focused on plotting data to identify features such as trends and seasonal variations. While this is a crucial first step, the next stage involves fitting time series models. We represent a time series of length n as $\{x_t : t = 1, \ldots, n\} = \{x_1, x_2, \ldots, x_n\}$, where n values are sampled at discrete times $t = 1, 2, \ldots, n$. When the series length is not essential, we abbreviate it as $\{x_t\}$.

A time series model is a sequence of random variables, and the observed series is a realization of this model. We use the same notation for both, with context distinguishing between them. An overline denotes sample means.

$$\bar{x} = \sum \frac{x_i}{n} \tag{1}$$

The 'hat' notation represents a prediction or forecast. For a series $\{x_t : t = 1, \ldots, n\}$, $\hat{x}_{t+k|t}$ denotes a forecast made at time t for the value at $t + k$. The number of steps into the future, k, is the lead time. Depending on the context, $\hat{x}_{t+k|t}$ may refer to either the random variable or its numerical value.

1.3.2 Models

Many time series are dominated by trend and/or seasonal effects. A simple additive decomposition model is given by:

$$x_t = m_t + s_t + z_t \tag{2}$$

where x_t is the observed series, m_t is the trend, s_t is the seasonal effect, and z_t is the error term, often a sequence of correlated random variables with mean zero. Two main approaches for extracting m_t and s_t will be outlined along with R functions for this.

For cases where the seasonal effect increases with the trend, a multiplicative model may be more suitable:

$$x_t = m_t \cdot s_t + z_t \tag{3}$$

Alternatively, an additive decomposition for $\log(x_t)$ can be used:

$$\log(x_t) = m_t + s_t + z_t \tag{4}$$

Care is needed when transforming back to x_t from $\log(x_t)$ to avoid bias. If z_t is normally distributed with mean 0 and variance σ^2, the predicted mean value is:

$$\hat{x}_t = e^{m_t + s_t + \frac{1}{2}\sigma^2} \tag{5}$$

For non-normal distributions, bias correction may lead to overcorrection, requiring an empirical adjustment. This is critical, for instance, in financial forecasts, where underestimating mean costs is a common issue.

1.3.3 Estimating trends and seasonal effects

A simple way to estimate the trend m_t is by calculating a moving average centered on x_t. A moving average smooths the time series by averaging a specified number of values around each x_t, except for the first and last few terms. For monthly data, the moving average spans 12 months. Since the average of $t = 1$ (January) to $t = 12$ (December) falls between June and July (i.e., $t = 6.5$), we average two consecutive moving averages to center the result at $t = 7$. The centered moving average for m_t is given by:

$$\hat{m}_t = \frac{1}{12}\left(\frac{1}{2}x_{t-6} + x_{t-5} + \cdots + x_{t+5} + \frac{1}{2}x_{t+6}\right) \tag{6}$$

where $t = 7, \ldots, n-6$. The coefficients sum to 1, ensuring equal weight for each value. This method generalizes to other seasonal frequencies (e.g., quarterly) by maintaining the condition that coefficients sum to unity.

The seasonal effect $\hat{s}_t$ can be estimated by subtracting the trend:

$$\hat{s}_t = x_t - \hat{m}_t \tag{7}$$

Averaging the monthly estimates across all years provides a single estimate of the effect for each month. To ensure the seasonal effects sum to zero, they are adjusted by subtracting the mean. For multiplicative models, the estimate becomes:

$$\hat{s}_t = \frac{x_t}{\hat{m}_t} \tag{8}$$

and multiplicative factors are adjusted to average to 1. Seasonally adjusted data, often used in economic indicators, removes seasonal effects. If the seasonal effect is additive, the adjusted series is $x_t - \bar{s}_t$, and if multiplicative, it is $x_t/\bar{s}_t$, where $\bar{s}_t$ is the mean seasonal adjustment for the given time.

1.3.4 Smoothing

The centred moving average is a smoothing procedure applied retrospectively to identify an underlying trend in a time series. It uses points before and after the target time, often leaving some missing values at the series' start and end unless adapted for edge points. Another smoothing method in R is 'stl', which uses locally weighted regression (loess). This local regression considers a small number of points around the target time, weighted to reduce the influence of outliers, making it a robust regression. While straightforward in principle, the details of 'stl' are complex.

Unlike smoothing, which does not provide a forecast model, fitting a linear trend has the advantage of enabling extrapolation. The term "filtering" is also used in this context, particularly in engineering, to describe obtaining the best estimate of a variable based on past and current noisy measurements. Filtering is vital in control algorithms, such as those used by the Huygens probe during its 2005 landing on Titan.

1.3.5 Decomposition in *R*

In R, the function 'decompose' estimates trends and seasonal effects using a moving average. Nesting it within 'plot' (e.g., 'plot(stl())') produces a figure showing the original series x_t, and decomposed series m_t, s_t, and z_t. For example, additive and multiplicative decomposition plots for electricity data are created by the following commands, with the seasonal effect superimposed on the trend using 'lty' for line types.

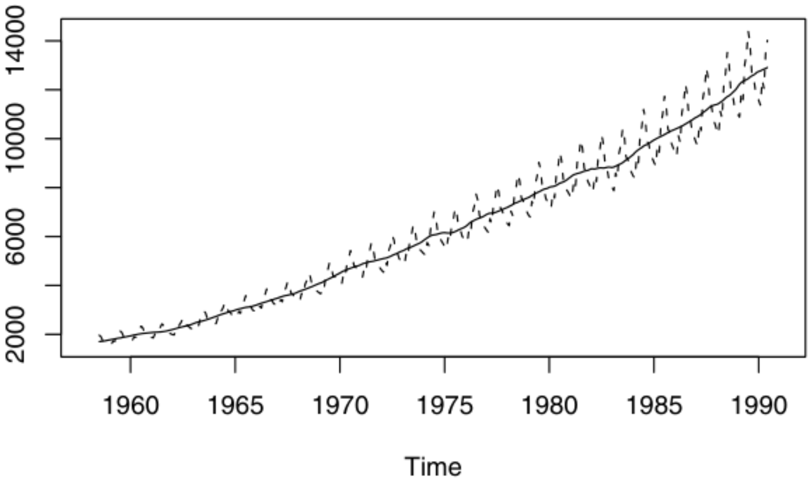


Figure 1: Electricity production data: trend with superimposed multiplicative seasonal effects.

```r
# Decomposition of the time series
plot(decompose(Elec.ts))

# Multiplicative decomposition
Elec.decom <- decompose(Elec.ts, type = "mult")
plot(Elec.decom)

# Extracting the trend and seasonal components
Trend <- Elec.decom$trend
Seasonal <- Elec.decom$seasonal

# Plotting the trend and the product of trend and seasonal effect
ts.plot(cbind(Trend, Trend * Seasonal), lty = 1:2)
```

Listing 1: Decomposition of Time Series in R

A multiplicative model is often more suitable than an additive one when the variance of the series and trend increase over time. However, if the random component z_t also shows increasing variance, a log-transformation (Eq. 1.4) may be more appropriate.

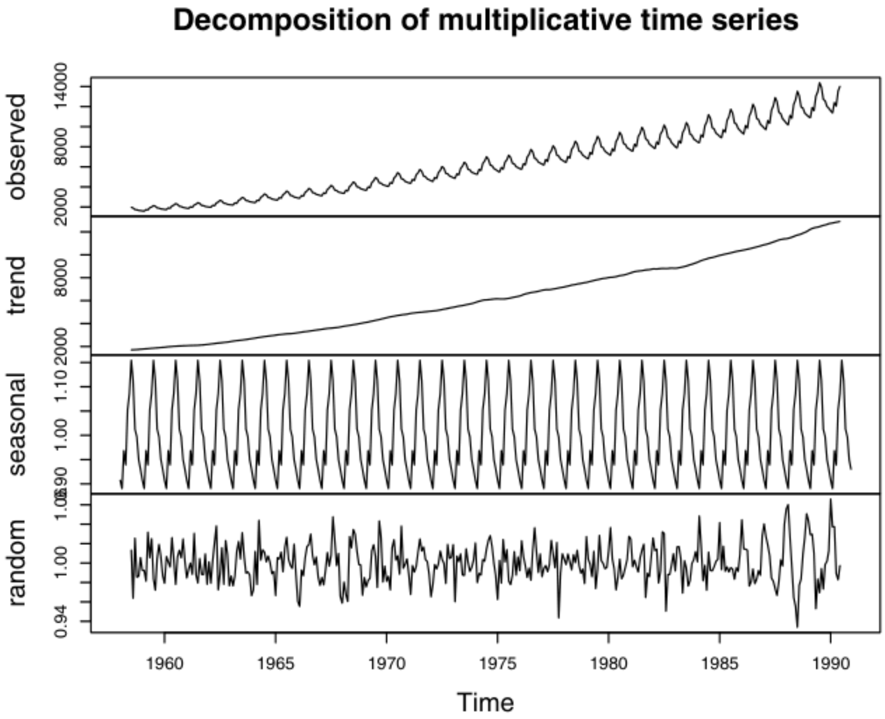


Figure 2: Decomposition of the electricity production data.

The random series from ‘decompose‘ is not the true realization of z_t, but an estimate derived from the trend and seasonal components, treated as a residual error series, yet used as a realisation of the random process.

Module - 1

Chapter - 2

2 Characteristics of Time Series

A time series is a collection of data points indexed in time order, usually spaced at uniform intervals. It captures observations at successive points in time, and this ordering makes time series distinct from other types of data because time series analysis is inherently about trends, patterns, and dependencies across time. To effectively analyze time series data, it is important to understand its core characteristics. These characteristics form the basis for any meaningful analysis, forecasting, or modeling.

2.1 Introduction and Examples

Time series is a sequence of data points collected or recorded at regular time intervals. It is used in various domains such as economics, finance, meteorology, medicine, and engineering to analyze trends, patterns, and seasonal variations. Unlike cross-sectional data, time series data captures the dynamics and changes over time, allowing for forecasting and insight extraction from historical patterns.

Common examples of time series data include:

- Stock market prices recorded every minute.
- Daily temperature recordings in a city.
- Monthly sales data for a retail store.
- Quarterly GDP of a country.
- Yearly rainfall data in a specific region.

In R, we can visualize a simple time series data using the `AirPassengers` dataset:

```r
# Example in R:
data("AirPassengers")
plot(AirPassengers, main="AirPassengers Dataset",
     ylab="Number of Passengers", xlab="Year")
```

Listing 2: Example in R

2.2 Objectives and Nature of Time Series

The main objectives of time series analysis include:

- **Understanding** the underlying patterns in the data.
- **Identifying** components such as trend, seasonality, and cyclic behaviors.
- **Modeling** the time series to predict future values.
- **Detecting anomalies** or unexpected changes.
- **Smoothing** to eliminate noise and reveal important patterns.

The nature of time series can be categorized into:

- **Trend Component (T)**: A long-term increase or decrease in the data. For example, the overall upward movement in stock market prices over several years.

- **Seasonal Component (S)**: Regular fluctuations that repeat over a specific period, such as monthly sales peaking every December.
- **Cyclic Component (C)**: Recurrent but non-periodic fluctuations often linked to economic cycles.
- **Irregular Component (I)**: Random variations that do not follow a pattern.

These components can be combined using the following additive or multiplicative models:

$$Y(t) = T(t) + S(t) + C(t) + I(t) \tag{9}$$

or

$$Y(t) = T(t) \times S(t) \times C(t) \times I(t) \tag{10}$$

In R, we can decompose a time series to analyze these components:

```r
# Decomposition Example in R:
decomposed <- decompose(AirPassengers)
plot(decomposed)
```

Listing 3: Decomposition Example in R

2.3 Introduction to Time Series Databases and Applications

Time series databases (TSDB) are optimized for storing and querying time series data. Unlike traditional relational databases, TSDBs are built to efficiently handle timestamped data, which makes them ideal for applications that require storing large volumes of time-dependent data with high write-throughput and quick retrieval.

Some popular time series databases include:

- **InfluxDB**: InfluxDB is an open-source time series database designed specifically for storing and managing time series data such as metrics, events, and analytics. It is widely used in the Internet of Things (IoT), DevOps, real-time analytics, and monitoring applications. InfluxDB is part of the InfluxData stack, which includes Telegraf (a plugin-based collector of metrics), Chronograf (a visualization tool), and Kapacitor (an alerting and processing tool).

  InfluxDB uses a simple and flexible query language called InfluxQL, which is similar to SQL but optimized for time series operations like aggregations, time windowing, and transformations. It also supports the newer Flux language, which offers more powerful queries. InfluxDB is known for its high write throughput, enabling it to handle millions of writes per second, which makes it ideal for real-time data collection and analysis.

  InfluxDB stores data in a compressed time series format and offers a retention policy feature, where users can automatically delete old data to manage storage effectively. It also supports downsampling, which helps reduce storage costs by keeping high-resolution data for shorter periods while retaining summarized, lower-resolution data for longer periods.

  InfluxDB uses the InfluxQL or Flux query languages. Below is an example using InfluxQL.

  ```sql
  # Storing time series data in InfluxDB
  INSERT temperature,location=room1 value=72.5 1627550220
  INSERT temperature,location=room2 value=75.3 1627550280
  ```

  Listing 4: Storing Time Series Data in InfluxDB

  ```sql
  # Querying time series data in InfluxDB
  SELECT mean("value") FROM "temperature"
  WHERE time >= '2021-07-29T00:00:00Z' AND time < '2021-07-30T00:00:00Z'
  GROUP BY time(1h), "location" fill(none)
  ```

  Listing 5: Querying Time Series Data in InfluxDB

  Advantages of InfluxDB:

  - **High Performance**: InfluxDB is optimized for fast writes, making it ideal for use cases where data is collected frequently, such as IoT applications or system monitoring. Its write throughput is among the best in class for time series databases.
  - **Retention Policies**: Users can define retention policies to automatically expire older data, thus managing storage costs efficiently. This is particularly useful in environments where data grows exponentially over time.
  - **Schema-Free**: InfluxDB is schemaless, meaning that data can be written with any fields and tags, making it flexible to adapt to new use cases and metrics without predefined structures.
  - **Integrations**: InfluxDB integrates easily with other tools such as Grafana for visualization, Telegraf for data collection, and Kapacitor for alerting and data processing.

  Cons of InfluxDB:

  - **Query Complexity**: While InfluxQL is simple for basic queries, more complex queries involving joins or transformations might be challenging. Flux, the newer query language, addresses these issues but introduces a learning curve.
  - **Scaling Issues**: Scaling InfluxDB horizontally (i.e., across multiple nodes) can be challenging. The enterprise version of InfluxDB offers clustering, but the open-source version does not, making scaling limited for high-availability deployments.
  - **Storage Costs**: Although InfluxDB offers compression, the storage requirements for high-frequency data can still be substantial, especially in long-term retention scenarios.

- **TimescaleDB**: TimescaleDB is an open-source time series database built as an extension to PostgreSQL. By leveraging the mature and robust PostgreSQL ecosystem, TimescaleDB inherits features like ACID compliance, powerful indexing, relational joins, and SQL support, making it a reliable option for time series applications that also require relational data.

  TimescaleDB splits large datasets into ”chunks” based on time intervals, which allows for efficient time-series queries. This method also makes TimescaleDB horizontally scalable while keeping storage costs down. Additionally, TimescaleDB supports native compression, significantly reducing the storage footprint for large datasets.

  One of the strengths of TimescaleDB is its seamless integration with the broader PostgreSQL ecosystem, which includes extensions, tools, and libraries. This makes it an attractive choice for users who already have a PostgreSQL setup and are looking to incorporate time series capabilities into their existing infrastructure without migrating to a new system.

  TimescaleDB uses standard SQL, with time series data stored in hypertables.

```sql
-- Creating a hypertable in TimescaleDB
CREATE TABLE temperature (
    time        TIMESTAMPTZ       NOT NULL,
    location    TEXT              NOT NULL,
    temperature DOUBLE PRECISION  NOT NULL
);

-- Convert the table to a hypertable
SELECT create_hypertable('temperature', 'time');

-- Inserting data into TimescaleDB
INSERT INTO temperature (time, location, temperature)
VALUES (NOW(), 'room1', 72.5), (NOW(), 'room2', 75.3);
```

Listing 6: Storing Time Series Data in TimescaleDB

```sql
-- Querying the average temperature by hour
SELECT time_bucket('1 hour', time) AS bucket, location, avg(temperature)
FROM temperature
WHERE time > NOW() - interval '24 hours'
GROUP BY bucket, location
```

```sql
ORDER BY bucket DESC;
```

Listing 7: Querying Time Series Data in TimescaleDB

Advantages of TimescaleDB:

- **PostgreSQL Ecosystem**: Since TimescaleDB is an extension of PostgreSQL, it benefits from the stability, reliability, and community support of PostgreSQL. This includes support for advanced indexing, relational joins, transactions, and other features common to relational databases.
- **SQL Support**: TimescaleDB supports standard SQL queries, making it easy for developers familiar with SQL to work with time series data. This reduces the learning curve compared to other TSDBs that use custom query languages.
- **Efficient Time Series Storage**: TimescaleDB automatically partitions data into chunks based on time intervals, which improves query performance. It also supports data compression, making it highly efficient for storing large datasets.
- **Scalability**: TimescaleDB provides built-in tools for scaling horizontally, allowing it to handle large time series datasets across distributed environments.

Cons of TimescaleDB:

- **Limited for Extreme Real-Time Use Cases**: While TimescaleDB performs well for most time series applications, it may not be as optimized for extreme high-frequency, real-time applications as InfluxDB or Prometheus.
- **Complexity with Large Joins**: Although relational joins are a strength of TimescaleDB, performing large-scale joins on massive datasets can lead to performance issues, particularly for real-time queries.
- **Enterprise Features**: Some advanced features, like continuous aggregation and advanced compression, are part of TimescaleDB's enterprise offering, which can be a limitation for users relying only on the open-source version.

- **Prometheus**: Prometheus is a highly popular, open-source monitoring and alerting toolkit designed specifically for cloud-native environments. It was developed as part of the Cloud Native Computing Foundation and is often used in conjunction with Kubernetes for monitoring application performance, infrastructure metrics, and other system behaviors.

  Prometheus works by scraping metrics from instrumented services at regular intervals, storing them as time series data. It supports multi-dimensional data collection using labels, which are key-value pairs attached to the metrics. Prometheus uses its own query language called PromQL (Prometheus Query Language), which is specifically designed for aggregating and filtering time series data. Its alerting mechanism is flexible and integrates easily with various notification systems like PagerDuty, Slack, and email.

  One of the primary use cases for Prometheus is in monitoring cloud infrastructure, where it excels at tracking the performance of servers, containers, and microservices. The system is designed to be lightweight and works well in environments where quick real-time insights and monitoring are critical.

  Prometheus uses PromQL for querying, and data is scraped from instrumented services.

```yaml
# An example of Prometheus scrape configuration
scrape_configs:
  - job_name: 'node_exporter'
    static_configs:
      - targets: ['localhost:9100']
```

Listing 8: Scraping Time Series Data in Prometheus

```
# Querying time series data in Prometheus using PromQL
avg_over_time(node_cpu_seconds_total[5m])
```

Listing 9: Querying Time Series Data in Prometheus

Advantages of Prometheus:

- **Cloud-Native Friendly**: Prometheus is designed to work seamlessly in dynamic cloud-native environments, particularly with Kubernetes. It is well-suited for containerized environments where services come and go frequently.
- **Highly Scalable**: Prometheus is built for large-scale, distributed environments. Its pull-based metrics collection makes it efficient for monitoring hundreds or thousands of services.
- **Alerting System**: Prometheus has a powerful alerting system that allows users to define alerting rules based on metric thresholds. It integrates easily with notification tools, enabling quick responses to system failures or abnormal metrics.
- **Multi-Dimensional Data Collection**: Prometheus allows users to attach labels to their metrics, making it easy to filter and aggregate data along different dimensions, such as by service, data center, or cluster.

Cons of Prometheus:

- **Limited Long-Term Storage**: Prometheus is not designed for long-term data retention. While it excels at real-time monitoring, users often need to integrate it with external databases like Thanos or Cortex to store time series data for long-term historical analysis.
- **No Built-In Clustering**: Prometheus does not support clustering in its native form, which can be a limitation for users requiring high availability and fault tolerance without external dependencies.
- **Query Language Complexity**: PromQL, the query language used by Prometheus, can be complex for new users, particularly for those used to SQL or other more common query languages. Learning to write efficient queries in PromQL can take time.

Applications of time series databases include:

- **IoT Data Storage**: Time series databases are commonly used in IoT devices to store sensor data, such as temperature readings, GPS data, or humidity levels.
- **Financial Market Analysis**: TSDBs handle high-frequency trading data, storing stock prices, trading volumes, and other financial indicators over time.
- **DevOps Monitoring**: Tracking system performance metrics like CPU usage, memory consumption, and network bandwidth usage in real-time.

An R example to work with a simple time series dataset:

```r
# Simulate time series data and store it
time_series_data <- ts(rnorm(100), frequency=12, start=c(2020, 1))
plot(time_series_data, main="Simulated Time Series Data",
     ylab="Values", xlab="Time")
```

Listing 10: Simulate Time Series Data in R

2.4 Measures of Dependence

2.4.1 Introduction to Measures of Dependence

In time series analysis, *measures of dependence* refer to the statistical relationships between observations in a time series dataset, especially over time lags. Understanding these dependencies is crucial because they indicate whether and how past values influence future values. A time series is dependent when the values at different points in time are not independent but rather exhibit a relationship that we can measure and analyze.

One of the primary measures of dependence in time series is the **autocorrelation function** (ACF). This function helps determine how observations at different time points are correlated with one another. The ACF for a time series $\{X_t\}$ at lag k is given by:

$$\rho(k) = \frac{\mathrm{Cov}(X_t, X_{t+k})}{\sqrt{\mathrm{Var}(X_t) \cdot \mathrm{Var}(X_{t+k})}}$$

where Cov denotes covariance, and Var represents the variance of the time series.

The partial autocorrelation function (PACF) is another measure of dependence, which describes the relationship between an observation and its lagged values, excluding the influence of intermediate lags. This is particularly useful in autoregressive (AR) models, where PACF helps determine the order of the model.

2.4.2 Example Problem: Estimating Autocorrelation

Consider a simple time series dataset where we observe daily stock prices for 10 days:

$$X = \{120, 122, 119, 118, 121, 123, 124, 125, 126, 128\}.$$

We want to calculate the autocorrelation at lag 1 and lag 2.

Step 1: Compute the mean of the series. First, compute the mean $\bar{X}$ of the series:

$$\bar{X} = \frac{120 + 122 + 119 + 118 + 121 + 123 + 124 + 125 + 126 + 128}{10} = \frac{1226}{10} = 122.6.$$

Step 2: Define the autocorrelation formula. The formula for autocorrelation at lag k is:

$$\rho(k) = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2},$$

where:

- X_t is the value at time t,
- $\bar{X}$ is the mean of the time series,
- k is the lag,
- n is the total number of observations.

Step 3: Calculate the denominator for all lags. The denominator for both lag 1 and lag 2 is the same:

$$\sum_{t=1}^{n}(X_t - \bar{X})^2 = (120 - 122.6)^2 + (122 - 122.6)^2 + (119 - 122.6)^2 + \ldots + (128 - 122.6)^2.$$

Substitute the values:

$$= (-2.6)^2 + (-0.6)^2 + (-3.6)^2 + (-4.6)^2 + (-1.6)^2 + (0.4)^2 + (1.4)^2 + (2.4)^2 + (3.4)^2 + (5.4)^2.$$

$$= 6.76 + 0.36 + 12.96 + 21.16 + 2.56 + 0.16 + 1.96 + 5.76 + 11.56 + 29.16 = 92.4.$$

Thus, the denominator is 92.4.

Step 4: Calculate the numerator for lag 1. Now, compute the numerator for lag 1, which is:

$$\sum_{t=1}^{n-1}(X_t - \bar{X})(X_{t+1} - \bar{X}) = (120 - 122.6)(122 - 122.6) + (122 - 122.6)(119 - 122.6) + \ldots + (126 - 122.6)(128 - 122.6).$$

Substitute the values:

$$= (-2.6)(-0.6) + (-0.6)(-3.6) + (-3.6)(-4.6) + (-4.6)(-1.6) + (-1.6)(0.4) + (0.4)(1.4) + (1.4)(2.4) + (2.4)(3.4) + (3.4)(5.4).$$

$$= 1.56 + 2.16 + 16.56 + 7.36 + (-0.64) + 0.56 + 3.36 + 8.16 + 18.36 = 57.68.$$

Step 5: Calculate autocorrelation at lag 1. Now that we have the numerator and denominator, calculate the autocorrelation:

$$\rho(1) = \frac{57.68}{92.4} \approx 0.624.$$

Thus, the autocorrelation at lag 1 is approximately 0.624, indicating a moderate positive correlation between consecutive values.

Step 6: Calculate the numerator for lag 2. For lag 2, compute the numerator:

$$\sum_{t=1}^{n-2}(X_t - \bar{X})(X_{t+2} - \bar{X}) = (120 - 122.6)(119 - 122.6) + (122 - 122.6)(118 - 122.6) + \ldots + (125 - 122.6)(128 - 122.6).$$

Substitute the values:

$$= (-2.6)(-3.6) + (-0.6)(-4.6) + (-3.6)(-1.6) + (-4.6)(0.4) + (-1.6)(1.4) + (0.4)(2.4) + (1.4)(3.4) + (2.4)(5.4).$$

$$= 9.36 + 2.76 + 5.76 + (-1.84) + (-2.24) + 0.96 + 4.76 + 12.96 = 32.48.$$

Step 7: Calculate autocorrelation at lag 2. Finally, calculate the autocorrelation at lag 2:

$$\rho(2) = \frac{32.48}{92.4} \approx 0.352.$$

Thus, the autocorrelation at lag 2 is approximately 0.352, indicating a weaker positive correlation between values that are two time steps apart.

Conclusion From these calculations, we see that the autocorrelation at lag 1 is higher (0.624) compared to lag 2 (0.352). This suggests that consecutive stock prices are more closely related than prices separated by two days, which is a typical observation in time series where immediate past values have a stronger influence on the present.

2.5 Stationary Time Series

2.5.1 Definition and Importance of Stationarity

A *stationary time series* is one whose statistical properties such as mean, variance, and autocorrelation are constant over time. Stationarity is important because many time series models, such as ARMA, ARIMA, or GARCH, assume that the data is stationary. If the data is not stationary, the model's performance can suffer, leading to poor forecasts.

Formally, a time series $\{X_t\}$ is stationary if for all time points t, the following conditions hold:

- $\mathbb{E}(X_t) = \mu$ (constant mean)
- $\mathrm{Var}(X_t) = \sigma^2$ (constant variance)
- $\mathrm{Cov}(X_t, X_{t+k}) = \gamma(k)$ (constant autocovariance that depends only on lag k)

2.5.2 Features of Stationary Time Series

Key characteristics of a stationary series include:

- The series fluctuates around a constant mean.
- There is no long-term trend.
- The autocorrelation function decreases quickly as the lag increases.

2.5.3 R Example: Plotting a Stationary Series

We can generate and plot a stationary time series in R using the following code:

```r
set.seed(123)
stationary_series <- ts(rnorm(100), frequency=12)
plot(stationary_series, main="Stationary Time Series", ylab="Values", xlab="Time")
```

Listing 11: Generating a Stationary Time Series

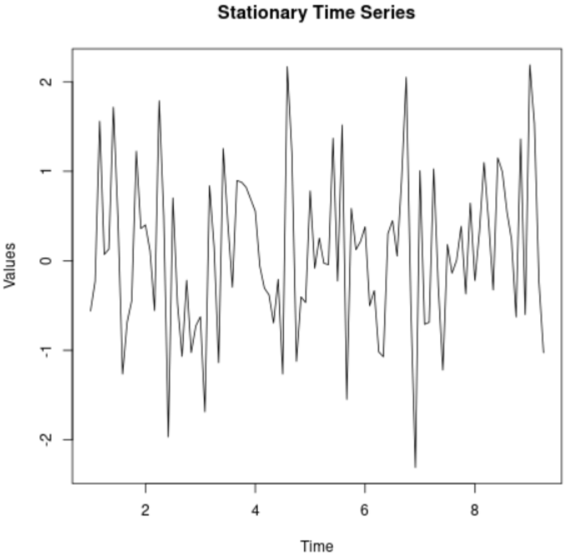


Figure 3: Stationary Time Series

The plot of the stationary series will show random fluctuations around a constant mean with no visible trend.

2.6 Estimation of Correlation

2.6.1 Definition of Correlation

Correlation is a measure of the strength and direction of the linear relationship between two variables. In the context of time series, correlation helps identify how strongly values at one time point relate to values at a later time point (lagged values). Estimating correlation is crucial for understanding the underlying patterns in the data, such as seasonality or trends.

The Pearson correlation coefficient is given by:

$$r = \frac{\sum_{t=1}^{n}(X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{n}(X_t - \bar{X})^2 \sum_{t=1}^{n}(Y_t - \bar{Y})^2}}$$

In time series analysis, we often compute *autocorrelation* or the correlation between values at different time lags. Estimating the autocorrelation function (ACF) helps us determine whether previous values have predictive power for future values.

2.6.2 Proof of Correlation for Time Series

For a time series $\{X_t\}$, the autocorrelation at lag k is:

$$\rho(k) = \frac{\sum_{t=k+1}^{n}(X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

This equation calculates the correlation between observations separated by k time steps. As k increases, $\rho(k)$ typically decreases, reflecting the diminishing influence of earlier observations on future values.

2.7 Vector-Valued and Multi-Dimensional Series

2.7.1 Definition and Importance

A *vector-valued time series* consists of multiple time series observed together. These can be considered multidimensional or multivariate, where each dimension represents a different but related time series. Analyzing such series is important in fields like economics (e.g., analyzing stock prices for multiple companies) and environmental science (e.g., temperature, humidity, and wind speed together).

Multidimensional time series analysis focuses on understanding the relationships between these multiple series and how they jointly evolve over time. The vector autoregressive (VAR) model is a common model used for such series.

2.7.2 Example: Vector-Valued Series in R

We can create and analyze a vector-valued time series in R using the following code:

```r
# Simulate two related time series
set.seed(123)
ts1 <- ts(rnorm(100), frequency=12)
ts2 <- ts(rnorm(100, mean=2), frequency=12)

# Combine them into a multivariate time series
multi_series <- ts(cbind(ts1, ts2), frequency=12)

# Plot the multivariate series
plot(multi_series, main="Vector-Valued Time Series", col=c("blue", "red"), lty=1:2)
legend("topright", legend=c("Series 1", "Series 2"), col=c("blue", "red"), lty=1:2)
```

Listing 12: Creating and Plotting a Multidimensional Time Series

Equation for Multivariate Model: In a multivariate time series model, each variable depends on its own past values and the past values of other variables. The vector autoregressive (VAR) model for two time series X_t and Y_t is given by:

$$X_t = \alpha_1 X_{t-1} + \beta_1 Y_{t-1} + \epsilon_{1t}$$
$$Y_t = \alpha_2 X_{t-1} + \beta_2 Y_{t-1} + \epsilon_{2t}$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are coefficients and $\epsilon_{1t}, \epsilon_{2t}$ are error terms.

This type of modeling is crucial in understanding how multiple series interact over time.

Module - 1

Chapter - 3 & 4

3 Components of Time Series

article amsmath

The characteristics of a time series are defined by the various types of movements or fluctuations that occur over time. These movements, known as the components of a time series, help explain the underlying patterns in the data. There are four main components:

1. Secular Trend (T)

The **Secular Trend**, also known as the **long-term trend** or simply **trend**, refers to the general movement of data, either upward or downward, over an extended period. It captures the long-term tendency of a dataset to grow or decline, ignoring short-term fluctuations.

For example, the population of India shows a clear upward trend over the years, while the death rate after independence has steadily declined due to improvements in literacy and healthcare. It's important to note that what constitutes a "long period" depends on the context of the data. For instance, an increase in cloth store sales over one year (e.g., from 1996 to 1997) is too short a period to be considered a secular trend.

However, in certain cases, a shorter time frame can reflect a trend if the nature of the data allows. For example, in a bacterial culture exposed to germicide, counting the number of organisms still alive every 10 seconds over 5 minutes could reveal a general decline in numbers, which would represent a secular trend over that period.

Mathematically, secular trends are categorized into two types:

1. **Linear Trend**: A consistent, straight-line increase or decrease over time.
2. **Curvi-Linear Trend (Non-Linear Trend)**: A trend where the rate of change is not constant, resulting in a curved pattern.

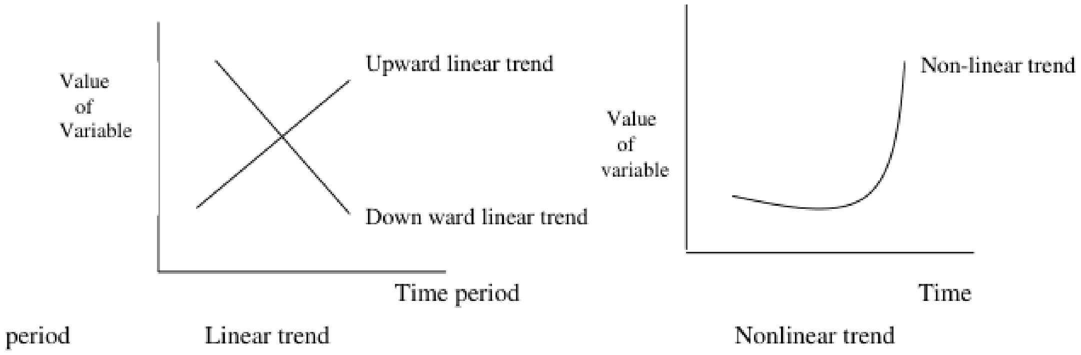


Figure 4: Linear Trend and Non Linear Trend

Example: A consistent rise in global temperatures over decades due to climate change.

2. Seasonal Variations (S)

Seasonal variations occur in a time series due to rhythmic forces that repeat in a regular and periodic manner within a period of less than one year. These variations follow the same pattern year after year. The period may be monthly, weekly, or even hourly, but if data is given in yearly terms, seasonal fluctuations do not exist.

Seasonal fluctuations in a time series arise from two main factors:

1. **Natural forces**
2. **Manmade conventions**

The most significant cause of seasonal variations is climate. Changes in weather conditions—such as rainfall, humidity, and temperature—impact industries and products differently. For example, there is a higher demand for woolen clothes and hot drinks in winter, while in summer, cotton clothes and cold drinks see increased sales. During the rainy season, the demand for umbrellas and raincoats rises.

In addition to nature, customs, traditions, and habits also influence seasonal variation. For instance, during festivals like Diwali, Dussehra, and Christmas, there is an increased demand for sweets and clothes. Similarly, the start of a school or college year sees a surge in demand for books and stationery.

Example: Higher sales of air conditioners during summer months due to the hot weather.

3. Cyclical Variations (C)

Cyclical movements occur over longer time periods than seasonal variations and typically reflect economic cycles such as booms and recessions. These cycles generally last for several years and, unlike seasonal variations, they do not follow a fixed or regular pattern.

Cyclical variations refer to short-term fluctuations lasting more than one year. The rhythmic movements in a time series that repeat in the same manner over a period longer than one year are called cyclical variations, and the duration is referred to as a **cycle**. Time series related to business and economics often exhibit cyclical behavior.

A classic example of cyclical variation is the **Business Cycle**, which includes four well-defined phases:

1. **Boom**: This phase is characterized by rapid economic growth, high levels of production, employment, and rising prices. During the boom period, consumer demand is strong, and businesses expand rapidly. However, inflationary pressures may also build up, leading to potential overheating of the economy.

   Example: The global economy in the late 1990s experienced a boom due to the dot-com bubble, where technology companies saw rapid growth and expansion.

2. **Decline**: After the boom, the economy begins to slow down. Production and demand decrease, unemployment starts to rise, and inflation stabilizes. This phase marks the transition from a peak towards a downturn, signaling the end of rapid economic expansion.

   Example: The early 2000s saw a decline after the burst of the dot-com bubble, where stock prices fell, and many tech companies collapsed, leading to a slowdown in economic growth.

3. **Depression**: This is the lowest phase of the cycle, marked by a significant decline in economic activity. There is high unemployment, reduced consumer spending, lower investment, and overall economic stagnation. It represents the most severe form of economic contraction.

   Example: The Great Depression of the 1930s is a classic example, where global economies shrank, unemployment reached record levels, and industrial output dropped sharply.

4. **Improvement (Recovery)**: After the depression, the economy begins to recover. Businesses start investing again, employment rises, and consumer confidence gradually returns. Production and demand start increasing, marking the beginning of the next upward cycle.

   Example: After the Great Recession of 2008, the economy began recovering in 2010, with improved job growth, increased consumer spending, and steady economic expansion.

These phases repeat over time, reflecting the fluctuating nature of economic activity.

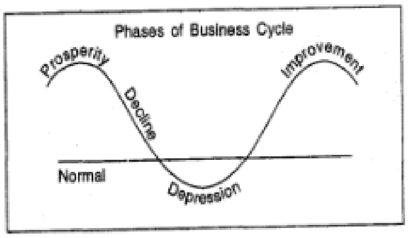


Figure 5: Phases of Business cycle

Example: Economic cycles with alternating periods of economic expansion and contraction.

4. Irregular Variations (I)

Irregular variations, also known as **Erratic**, **Accidental**, or **Random Variations**, are unpredictable and non-recurring fluctuations in a time series caused by unexpected events. Unlike trend, seasonal, and cyclical variations—which are considered regular variations—irregular variations are random and typically short-term, making them difficult to model or forecast.

These fluctuations are the result of unforeseen circumstances that are beyond human control, such as natural disasters, wars, pandemics, or other catastrophic events. Irregular variations significantly disrupt a time series but are not as structurally important as other variations.

Example: The COVID-19 pandemic in 2020 led to severe and unexpected disruptions across global economies, causing irregular variations in many time series related to employment, GDP, and stock market performance. This variation could not be predicted and doesn't follow any consistent or repeating pattern.

Together, these components provide a framework to analyze time series data, enabling better forecasting and understanding of the underlying patterns.

3.1 Additive and Multiplicative models

In time series analysis, a **mathematical model** represents the underlying structure of the data. It is assumed that the time series consists of various components such as trends, seasonal variations, cyclical variations, and irregular variations. These components together explain the observed value of the time series at any point in time.

The objective of a mathematical model is to decompose a time series into its constituent components in order to better understand, analyze, and forecast future values. Two widely used models in classical time series analysis are the **Additive Model** and the **Multiplicative Model**.

Why Are Mathematical Models Needed?

Mathematical models are essential in time series analysis for the following reasons:

1. **Understanding patterns**: By decomposing the time series into its components, we can identify trends, seasonal behaviors, and cyclical movements that help in understanding the nature of the data.

2. **Forecasting**: Mathematical models help us forecast future values based on past observations and the relationships among the components.

3. **Analyzing irregularities**: With the model, irregular or random variations can be separated from systematic variations, allowing analysts to focus on predictable aspects of the data.

Additive Model

The **Additive Model** assumes that the different components of a time series combine in an additive manner. That is, the observed value Y_t at time t is the sum of the contributions of the individual components.

Mathematically, the additive model is represented as:

$$Y_t = T_t + S_t + C_t + I_t$$

where:

- Y_t is the observed value at time t,
- T_t is the **trend component** at time t,
- S_t is the **seasonal component** at time t,
- C_t is the **cyclical component** at time t, and
- I_t is the **irregular component** at time t.

Derivation of Additive Model

The additive model is useful when the variation in the seasonal and cyclical components remains relatively constant over time. For example, if sales of ice cream increase by a fixed amount every summer, we can model that seasonal variation additively.

Example

Consider a time series of monthly sales data for a store over a year. Suppose the trend increases by 5 units per month, the seasonal effect adds 10 units during the summer months (June, July, and August), and cyclical factors add or subtract up to 3 units. Then, using the additive model, we can express the sales data Y_t for a summer month as:

$$Y_t = 5t + 10 + C_t + I_t$$

where C_t is the cyclical effect and I_t represents any irregular variations.

Multiplicative Model

The **Multiplicative Model** assumes that the components of the time series interact in a multiplicative manner. That is, the observed value Y_t at time t is the product of the contributions of the individual components.

Mathematically, the multiplicative model is represented as:

$$Y_t = T_t \times S_t \times C_t \times I_t$$

where the variables Y_t, T_t, S_t, C_t, and I_t represent the same components as in the additive model.

Derivation of Multiplicative Model

The multiplicative model is useful when the seasonal and cyclical variations are proportional to the level of the trend. For instance, if sales of ice cream double in the summer but are still dependent on an overall increasing trend, a multiplicative model would be more appropriate.

Example

Consider a company's quarterly revenue over a few years. If the trend increases by 10% each quarter, and sales are doubled during the holiday season, the multiplicative model expresses the revenue as:

$$Y_t = T_t \times 2 \times C_t \times I_t$$

where T_t represents the 10% growth in each quarter, the factor 2 accounts for the seasonal holiday surge, C_t captures any cyclical effects, and I_t represents irregular variations.

Choosing Between Additive and Multiplicative Models

The choice between the additive and multiplicative models depends on the nature of the data:

- **Additive Model**: Appropriate when the variations are constant over time and do not depend on the trend.
- **Multiplicative Model**: Suitable when the variations grow or shrink in proportion to the trend.

Conclusion

Both the additive and multiplicative models provide valuable ways to decompose a time series into its underlying components. By choosing the right model, analysts can gain better insights into trends, seasonal variations, and cyclical movements, and make more accurate forecasts.

3.2 Resolving components of a Time Series

In time series analysis, resolving the different components is a fundamental task to understand the underlying patterns. Time series data is usually composed of several components, and the key components include:

- **Trend** (T_t): The long-term movement in the data over time.
- **Seasonality** (S_t): Regular patterns that repeat over fixed intervals of time.
- **Cyclicality** (C_t): Long-term fluctuations caused by economic cycles.
- **Irregularity** (I_t): Random or unpredictable movements, typically caused by unforeseen factors.

The relationship between these components can be expressed using two main models:

- **Additive Model:**
$$Y_t = T_t + S_t + C_t + I_t$$
- **Multiplicative Model:**
$$Y_t = T_t \times S_t \times C_t \times I_t$$

In R, you can resolve components of a time series using built-in functions like ‘decompose()‘ for additive models or ‘stl()‘ for both additive and multiplicative models. Consider the following example where we decompose the AirPassengers dataset.

~~~
```r
# Load AirPassengers dataset
data(AirPassengers)

# Decompose the time series
decomposed_data <- decompose(AirPassengers, type = "multiplicative")

# Plot decomposed components
plot(decomposed_data)
~~~

Listing 13: Creating and Plotting a Multidimensional Time Series

3.3 Measuring Trend

The trend component of a time series reflects the long-term movement in the data. Understanding the trend is crucial for forecasting future values and identifying underlying patterns. There are several methods commonly used to measure trends:

3.3.1 Graphic

The graphic method, also known as the eye inspection method, is the simplest and most intuitive approach to identifying trends in time series data. This method involves the following steps:

1. **Plot the Data:** First, plot the given time series data on a graph, with time on the x-axis and the variable of interest on the y-axis.
2. **Draw a Trend Line:** A smooth, free-hand curve is then drawn through the plotted points, representing the general tendency of the series. This curve visually highlights the trend over time.

The graphic method effectively removes short-term variations to reveal the underlying trend in the data. The trend line can also be extended to predict or estimate future values, making it a useful tool for forecasting.

Importance of the Graphic Method

- Provides a visual and intuitive understanding of the trend.
- Easy to implement, requiring no complex calculations.
- Serves as a preliminary tool before applying more sophisticated methods.

Limitations

However, it is important to note that this method is subjective, and the accuracy of the predictions may vary depending on how the trend line is drawn. As such, while the graphic method is useful for initial analysis, it should be supplemented with more rigorous statistical techniques for reliable forecasting.

Example

Consider monthly sales data for a retail store over a year:

Month	Sales
Jan	100
Feb	120
Mar	140
Apr	160
May	150
Jun	130
Jul	180
Aug	190
Sep	170
Oct	160
Nov	140
Dec	200

Table 2: Monthly Sales Data

In R, the plotting can be done using the following code:

```r
# Define the sales data
sales <- c(100, 120, 140, 160, 150, 130, 180, 190, 170, 160, 140, 200)
months <- 1:12

# Plot the sales data
plot(months, sales, type = "o", col = "blue", xlab = "Month", ylab = "Sales")

# Add a manually drawn trend line (approximate)
lines(c(1, 12), c(100, 200), col = "red", lwd = 2)
```

Listing 14: Creating and Plotting a Multidimensional Time Series

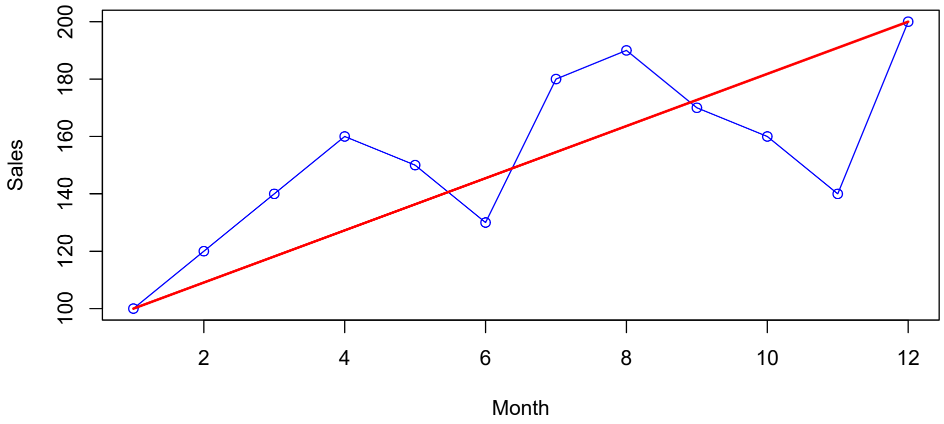


Figure 6: Trend from the Data

The red line represents the overall trend in sales. Although the data fluctuates, the general upward direction is clearly visible.

Advantages:

- **Simplicity:** The graphic method is one of the simplest approaches to studying trend values and is easy to implement.
- **Expertise Benefits:** An experienced statistician can often draw a trend line that better represents the data than one fitted using mathematical formulas.
- **Applicability:** Despite not being recommended for beginners, this method has significant merits in the hands of skilled statisticians and is widely used in practical applications.

Disadvantages:

- **Subjectivity:** The method is highly subjective; the resulting trend line can vary significantly based on who draws it.
- **Skill Requirements:** It requires the work to be conducted by skilled and experienced individuals to ensure accuracy.
- **Reliability Concerns:** The subjective nature of this method means that predictions derived from it may not be reliable.
- **Careful Execution:** Drawing the trend line must be done carefully to avoid misrepresentation of the data.

3.3.2 Semi-Averages

The semi-averages method involves dividing the time series data into two equal parts with respect to time. For instance, if we have data spanning from 1999 to 2016 (a total of 18 years), we would split it into two equal parts:

- The first part: 1999 to 2007 - The second part: 2008 to 2016

In cases where the number of years is odd, such as 9, 13, or 17, the middle year is omitted. For example, for 19 years of data from 1998 to 2016, the division would be:

- The first part: 1998 to 2006 - The second part: 2008 to 2016 (omitting the middle year 2007)

Once the data is divided, we calculate the arithmetic mean for each part, yielding two average values. These averages are then plotted against the mid-year of each part, and a straight line is drawn to connect the two points. This line represents the trend, which can be extended to estimate intermediate values or predict future values.

3.3.3 Example

Consider the following production data over several years:

Year	Production
2001	40
2002	45
2003	40
2004	42
2005	46
2006	52
2007	56
2008	61

Table 3: Production Data

To calculate the semi-averages:

1. **Divide the data:** - First part (2001 to 2004): 40, 45, 40, 42 - Second part (2005 to 2008): 46, 52, 56, 61

2. **Calculate the averages:** - First part average:

$$\text{Average}_1 = \frac{40 + 45 + 40 + 42}{4} = \frac{167}{4} = 41.75$$

- Second part average:

$$\text{Average}_2 = \frac{46 + 52 + 56 + 61}{4} = \frac{215}{4} = 53.75$$

3. **Plotting:**

- The averages (41.75 and 53.75) are plotted against the mid-years (2002.5 for the first part and 2006.5 for the second part). - A straight line is drawn connecting these two points, which represents the trend in production.

This method effectively captures the underlying trend in the data, providing a straightforward approach to trend analysis. The blue points represent the production data over the years, while the red points indicate the semi-averages calculated for the two parts. The red line shows the trend derived from these semi-averages.

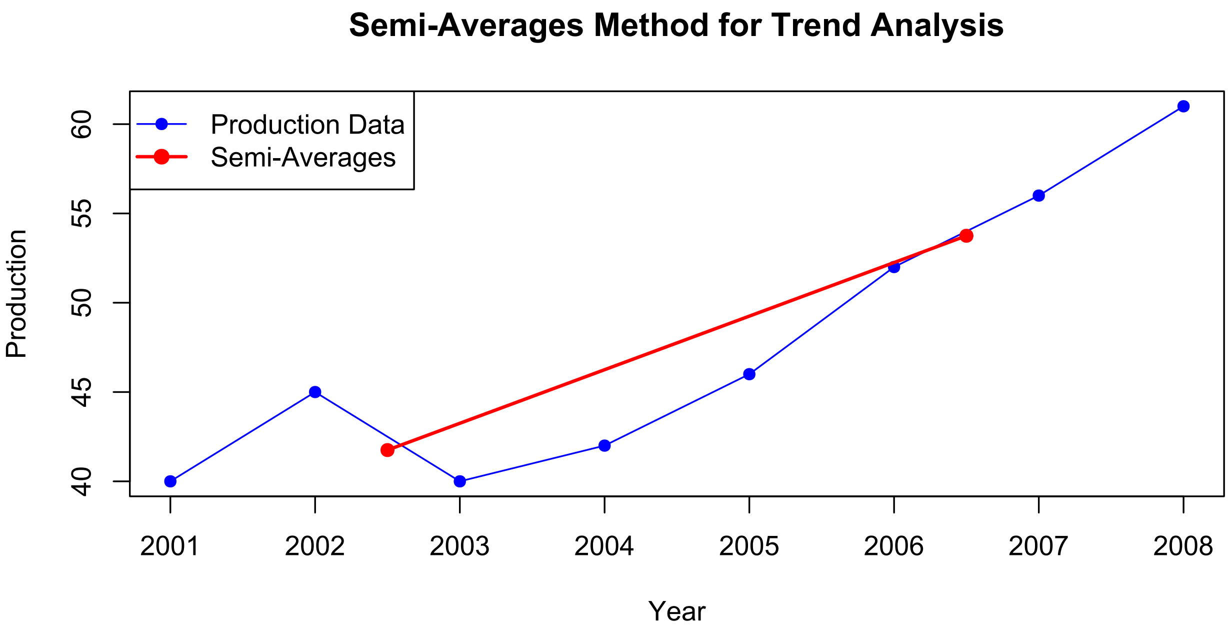


Figure 7: Trend analysis using the semi-averages method.

Advantages:

- **Simplicity:** This method is easier to understand compared to the moving average method and the method of least squares.
- **Objectivity:** It is an objective method for measuring trends; anyone applying this method will arrive at the same results.

Disadvantages:

- **Assumption of Linearity:** The method assumes a straight-line relationship between the plotted points, regardless of whether such a relationship actually exists.
- **Data Sensitivity:** If additional data is added to the original dataset, the entire calculation must be redone to obtain new trend values, and the trend line will change accordingly.
- **Influence of Extremes:** Since the arithmetic mean is calculated for each half, an extreme value in either half can significantly impact the points. As a result, the trend derived from these points may not be sufficiently accurate for future forecasting.

3.3.4 Moving Average

The moving average method is a widely used technique for computing trend values in a time series. This method effectively eliminates short-term and random fluctuations by calculating successive arithmetic means over a specified period. The period of the moving average is denoted as m, where m represents the number of data points included in each average.

The moving average is calculated as follows:

- The first average is the mean of the first m terms.
- The second average is the mean of the 2nd term to the $(m + 1)$th term.
- The third average is the mean of the 3rd term to the $(m + 2)$th term, and so on.

When m is odd, the moving average is associated with the mid-value of the time interval it covers. For instance, if $m = 3$, the moving average for the first three data points will be placed against the second data point (mid-point). However, if m is even, the moving average will lie between two middle periods, which do not correspond to any specific time period. To address this, a secondary calculation is performed by taking the average of the moving averages (2-yearly moving average) to align the result with a specific time period.

Example: Calculate the 3-yearly moving average for the following data.

Years	Production	3-Yearly Moving Average (Trend Values)
2001-02	40	
2002-03	45	$\frac{40+45+40}{3} = 41.67$
2003-04	40	$\frac{45+40+42}{3} = 42.33$
2004-05	42	$\frac{40+42+46}{3} = 42.67$
2005-06	46	$\frac{42+46+52}{3} = 46.67$
2006-07	52	$\frac{46+52+56}{3} = 51.33$
2007-08	56	$\frac{52+56+61}{3} = 56.33$
2008-09	61	

Table 4: 3-Yearly Moving Average Calculation

Calculation Explanation: - For 2002-03, the moving average is calculated using the production values for 2001-02, 2002-03, and 2003-04:

$$\text{Moving Average} = \frac{40+45+40}{3} = 41.67$$

- For 2003-04, the moving average uses the values for 2002-03, 2003-04, and 2004-05:

$$\text{Moving Average} = \frac{45+40+42}{3} = 42.33$$

This process continues until the last available data point. The moving average method is useful for smoothing out short-term fluctuations in data, providing a clearer view of the long-term trend. By systematically averaging data over a specified period, this method facilitates better forecasting and analysis in various fields, including economics, sales, and environmental studies.

Conclusion: The moving average is a fundamental tool in time series analysis, allowing for a better understanding of underlying trends by reducing noise from random fluctuations.

Calculate the 4-yearly moving average for the following data.

Years	Production	4-Yearly Moving Average	2-Yearly Moving Average (Trend Values)
2001-02	40		
2002-03	45		
2003-04	40	$\frac{40+45+40+42}{4} = 41.75$	$\frac{40+45}{2} = 42.5$
2004-05	42	$\frac{45+40+42+46}{4} = 43.15$	$\frac{40+42}{2} = 41$
2005-06	46	$\frac{40+42+46+52}{4} = 45$	$\frac{42+46}{2} = 44$
2006-07	52	$\frac{42+46+52+56}{4} = 49$	$\frac{46+52}{2} = 49$
2007-08	56	$\frac{46+52+56+61}{4} = 53.75$	$\frac{52+56}{2} = 54$
2008-09	61		

Table 5: 4-Yearly Moving Average Calculation

Calculation Explanation: - For 2003-04, the 4-yearly moving average is calculated as follows:

$$\text{Moving Average} = \frac{40+45+40+42}{4} = 41.75$$

- For 2004-05, the calculation is:

$$\text{Moving Average} = \frac{45+40+42+46}{4} = 43.15$$

- This process continues until the last available data point.
Additional Exercise Problems:

1. Given the following production data over a 5-year period, calculate the 3-yearly moving average. Use the moving averages to identify the trend:

   - 2010: 30
   - 2011: 35
   - 2012: 50
   - 2013: 45
   - 2014: 60

2. Consider the following data for sales over 6 years. Calculate the 2-yearly moving average and discuss any observed trends:

   - 2015: 80
   - 2016: 90
   - 2017: 85
   - 2018: 95
   - 2019: 100
   - 2020: 110

3. A company's quarterly earnings over two years are as follows. Calculate the 4-quarter moving average and explain any patterns you find:

   - Q1 2018: 200
   - Q2 2018: 220
   - Q3 2018: 210
   - Q4 2018: 250
   - Q1 2019: 240
   - Q2 2019: 260
   - Q3 2019: 280
   - Q4 2019: 300

Advantages:

- This method is simple to understand and easy to execute.
- It has flexibility in application; if new data for additional time periods are added, previous calculations remain unaffected, allowing for the generation of more trend values.
- It provides an accurate representation of the long-term trend, particularly if the trend is linear.
- When the period of the moving average coincides with the period of oscillation (cycle), periodic fluctuations are effectively eliminated.
- The moving average adapts to general movements in the data, with its shape determined by the actual data rather than arbitrary choices made by the statistician.
- It is effective for smoothing out short-term fluctuations, allowing for clearer visibility of long-term trends.
- The moving average can be easily visualized on a graph, making it a useful tool for presentations and reports.

Disadvantages:

- For a moving average of $2m + 1$, no trend values are generated for the first m and last m periods, limiting the analysis of the entire dataset.
- The trend path does not correspond to any specific mathematical function, making it unsuitable for forecasting or predicting future values.
- If the underlying trend is not linear, moving averages may not accurately reflect the true tendency of the data.
- The selection of the period for the moving average can be subjective, potentially introducing human bias into the analysis.
- Moving averages can lag behind actual data changes, which may lead to delays in identifying trends.
- In cases of sudden shifts or changes in the data, moving averages may provide a misleading representation of the trend, as they are based on historical data.
- The smoothing effect of moving averages can sometimes obscure important fluctuations that may need to be addressed.

3.3.5 Method of Least Squares

This method is widely used in practice. It is a mathematical approach that fits a trend line to the data, satisfying the following two conditions:

1. $\sum(Y - \hat{Y}) = 0$
2. $\sum(Y - \hat{Y})^2$ is minimized.

The method of least squares relies on two fundamental conditions to ensure that the fitted line provides the best representation of the data.

1. Condition: $P(Y - \hat{Y}) = 0$

This condition states that the sum of the residuals (the differences between the observed values Y and the predicted values $\hat{Y}$) must equal zero.

Explanation

- **Residuals**: The residual for each data point is defined as $Y_t - \hat{Y}_t$. It measures the error between the actual observation and the value predicted by the model.
- **Sum of Residuals**: When we sum these residuals across all observations, the condition $P(Y - \hat{Y}) = 0$ ensures that the positive and negative errors balance out. If this condition is satisfied, it indicates that the model does not systematically overestimate or underestimate the values.
- **Mathematical Justification**:
$$\sum(Y_t - \hat{Y}_t) = 0$$
This can also be derived from the optimization process, where minimizing the sum of squared deviations inherently leads to this condition.

2. Condition: $P(Y - \hat{Y})^2$ **is minimized**

This condition involves minimizing the sum of the squares of the residuals.

Explanation

- **Purpose of Squaring**: Squaring the residuals ensures that positive and negative errors do not cancel each other out, which could happen in the first condition. Squaring amplifies larger errors more than smaller ones, which helps in identifying models that fit better overall.
- **Objective**: The objective of the least squares method is to find the parameters (like a and b in a linear equation) that minimize the sum of these squared differences:
$$S = \sum(Y_t - \hat{Y}_t)^2$$

- **Geometric Interpretation**: In a geometric sense, this condition ensures that the trend line is as close as possible to all data points, minimizing the overall distance from each point to the line.
- **Derivation**: To find the best fitting line, we take the derivative of S with respect to the parameters (like a and b) and set these derivatives to zero. This process yields the normal equations, which can then be solved to find the optimal values of the parameters.

Summary
Both conditions together ensure that the best-fitting line through the data not only balances the residuals (no systematic bias) but also minimizes the overall error in terms of squared differences, leading to the most accurate predictions possible within the context of a linear model. This approach is foundational in regression analysis, helping to create models that accurately reflect underlying trends in data.

Fitting a Straight Line Trend by the Method of Least Squares Let Y_t be the value of the time series at time t. Thus, Y_t is the independent variable depending on t.
Assume a straight line trend of the form:

$$\hat{Y_t} = a + bt$$

where $\hat{Y_t}$ designates the trend values to distinguish them from the actual Y_t values, a is the Y-intercept, and b is the slope of the trend line.

To fit a straight line trend to a time series, we assume a linear relationship of the form:

$$Y_t^c = a + bt$$

where Y_t^c is the trend value at time t, a is the Y-intercept, and b is the slope of the trend line. The goal is to estimate the parameters a and b such that the sum of the squared deviations between the actual values Y_t and the trend values Y_t^c is minimized:

$$S = \sum (Y_t - Y_t^c)^2 = \sum (Y_t - (a + bt))^2.$$

To find the optimal values of a and b, we differentiate S with respect to a and b and set the derivatives to zero.

Differentiating S with respect to a gives:

$$\frac{\partial S}{\partial a} = -2 \sum (Y_t - (a + bt)) = 0.$$

Rearranging yields:

$$\sum (Y_t - (a + bt)) = 0 \implies \sum Y_t = na + b \sum t,$$

where n is the number of observations. Thus, we obtain:

$$na = \sum Y_t - b \sum t \implies a = \frac{1}{n} \left(\sum Y_t - b \sum t \right).$$

Substituting this back into the equation of Y_t^c leads to a simplified expression for a.

Next, we differentiate S with respect to b:

$$\frac{\partial S}{\partial b} = -2 \sum t (Y_t - (a + bt)) = 0.$$

Rearranging gives:

$$\sum t (Y_t - (a + bt)) = 0 \implies \sum t Y_t = a \sum t + b \sum t^2.$$

This can be rearranged to yield:

$$b = \frac{\sum t Y_t - a \sum t}{\sum t^2}.$$

The resulting normal equations from this process are:

$$(1) \quad \sum Y_t = na + b \sum t, \tag{11}$$

$$(2) \quad \sum t Y_t = a \sum t + b \sum t^2. \tag{12}$$

Solving these two normal equations will yield the estimates $\hat{a}$ and $\hat{b}$.

If we wish to fit a parabolic trend of the form:

$$Y_t^c = a + bt + ct^2,$$

we differentiate S with respect to c as well:

$$\frac{\partial S}{\partial c} = -2\sum(Y_t - (a + bt + ct^2))t^2 = 0.$$

Rearranging yields:

$$\sum(Y_t - (a + bt + ct^2))t^2 = 0 \implies \sum t^2 Y_t = a\sum t^2 + b\sum t^3 + c\sum t^4.$$

The normal equations for the parabolic trend can be summarized as:

$$(1) \quad \sum Y_t = na + b\sum t + c\sum t^2, \tag{13}$$

$$(2) \quad \sum tY_t = a\sum t + b\sum t^2 + c\sum t^3, \tag{14}$$

$$(3) \quad \sum t^2 Y_t = a\sum t^2 + b\sum t^3 + c\sum t^4. \tag{15}$$

Solving these three equations provides the values of $\hat{a}$, $\hat{b}$, and $\hat{c}$. Substituting these values into the equation for the parabolic trend gives:

$$Y_t^c = \hat{a} + \hat{b}t + \hat{c}t^2.$$

To assess the appropriateness of the parabolic trend model, one can use the method of second differences. If the second differences are constant (or nearly constant), the quadratic equation is a suitable representation of the trend component.

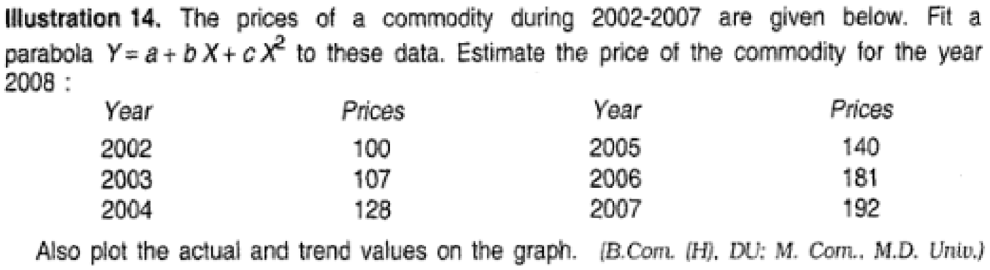
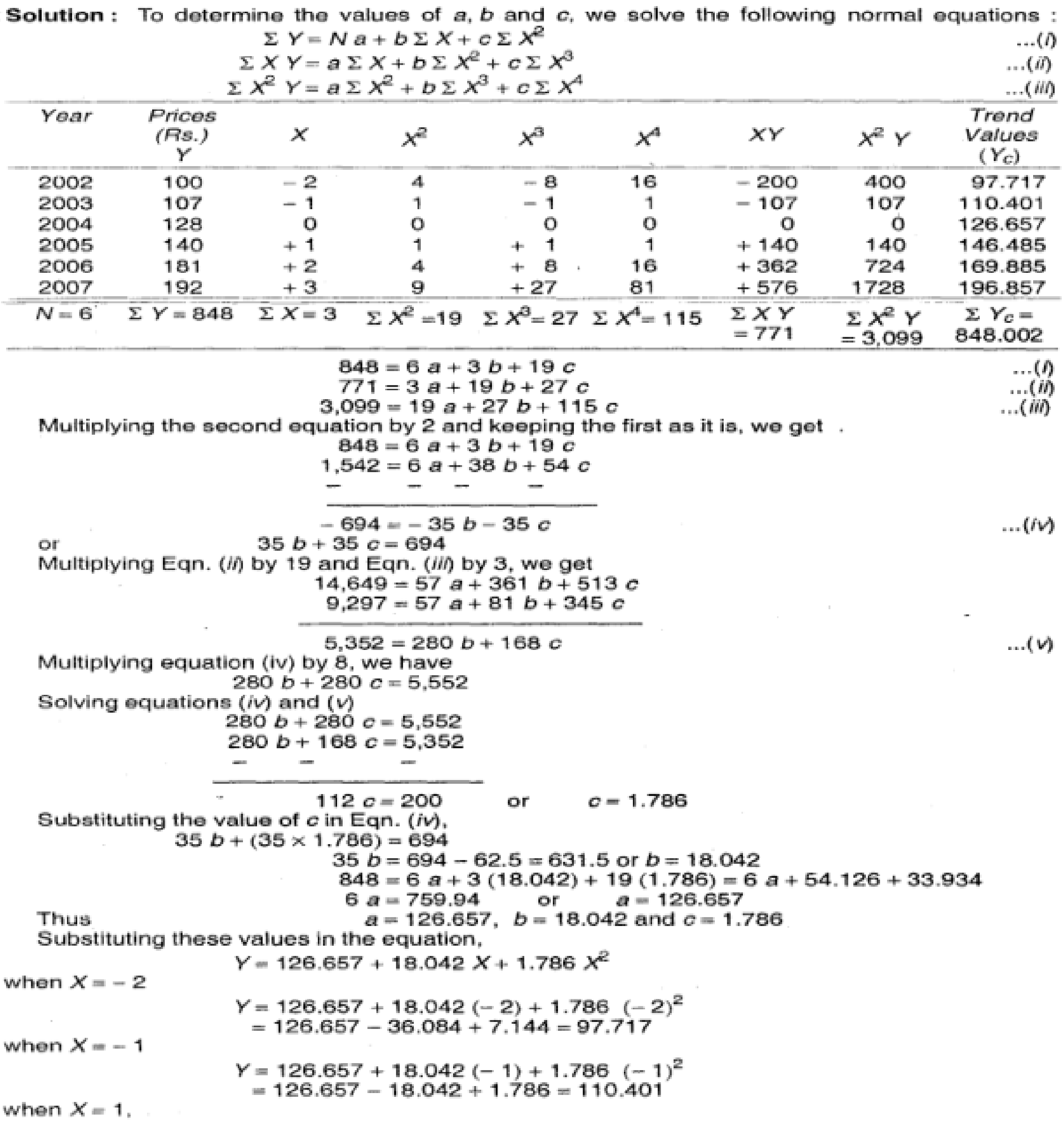


Illustration 14. The prices of a commodity during 2002-2007 are given below. Fit a parabola $Y = a + bX + cX^2$ to these data. Estimate the price of the commodity for the year 2008 :

Year	*Prices*	*Year*	*Prices*
2002	100	2005	140
2003	107	2006	181
2004	128	2007	192

Also plot the actual and trend values on the graph. *(B.Com. (H), DU: M. Com., M.D. Univ.)*



Solution : To determine the values of *a*, *b* and *c*, we solve the following normal equations :

$$\Sigma Y = Na + b\Sigma X + c\Sigma X^2 \quad \text{...}(i)$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3 \quad \text{...}(ii)$$
$$\Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 \quad \text{...}(iii)$$

Year	*Prices (Rs.) Y*	*X*	X^2	X^3	X^4	*XY*	X^2Y	*Trend Values* (Y_c)
2002	100	− 2	4	− 8	16	− 200	400	97.717
2003	107	− 1	1	− 1	1	− 107	107	110.401
2004	128	0	0	0	0	0	0	126.657
2005	140	+ 1	1	+ 1	1	+ 140	140	146.485
2006	181	+ 2	4	+ 8	16	+ 362	724	169.885
2007	192	+ 3	9	+ 27	81	+ 576	1728	196.857
$N = 6$	$\Sigma Y = 848$	$\Sigma X = 3$	$\Sigma X^2 = 19$	$\Sigma X^3 = 27$	$\Sigma X^4 = 115$	$\Sigma XY = 771$	$\Sigma X^2Y = 3{,}099$	$\Sigma Y_c = 848.002$

$$848 = 6a + 3b + 19c \quad \text{...}(i)$$
$$771 = 3a + 19b + 27c \quad \text{...}(ii)$$
$$3{,}099 = 19a + 27b + 115c \quad \text{...}(iii)$$

Multiplying the second equation by 2 and keeping the first as it is, we get .

$$848 = 6a + 3b + 19c$$
$$1{,}542 = 6a + 38b + 54c$$
$$- \quad - \quad - \quad -$$
$$-694 = -35b - 35c \quad \text{...}(iv)$$

or $\quad 35b + 35c = 694$

Multiplying Eqn. (*ii*) by 19 and Eqn. (*iii*) by 3, we get

$$14{,}649 = 57a + 361b + 513c$$
$$9{,}297 = 57a + 81b + 345c$$
$$5{,}352 = 280b + 168c \quad \text{...}(v)$$

Multiplying equation (iv) by 8, we have

$$280b + 280c = 5{,}552$$

Solving equations (*iv*) and (*v*)

$$280b + 280c = 5{,}552$$
$$280b + 168c = 5{,}352$$
$$- \quad - \quad -$$
$$112c = 200 \quad \text{or} \quad c = 1.786$$

Substituting the value of *c* in Eqn. (*iv*),

$$35b + (35 \times 1.786) = 694$$
$$35b = 694 - 62.5 = 631.5 \text{ or } b = 18.042$$
$$848 = 6a + 3(18.042) + 19(1.786) = 6a + 54.126 + 33.934$$
$$6a = 759.94 \quad \text{or} \quad a = 126.657$$

Thus $\quad a = 126.657, \quad b = 18.042$ and $c = 1.786$

Substituting these values in the equation,

$$Y = 126.657 + 18.042X + 1.786X^2$$

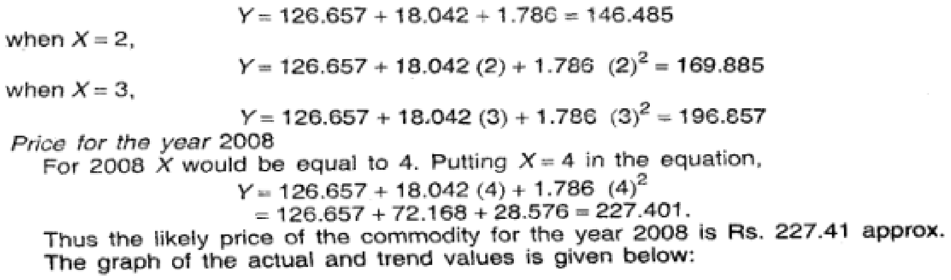
when $X = -2$

$$Y = 126.657 + 18.042(-2) + 1.786\ (-2)^2$$
$$= 126.657 - 36.084 + 7.144 = 97.717$$

when $X = -1$

$$Y = 126.657 + 18.042(-1) + 1.786\ (-1)^2$$
$$= 126.657 - 18.042 + 1.786 = 110.401$$

when $X = 1$,



$Y = 126.657 + 18.042 + 1.786 = 146.485$

when $X = 2$,

$Y = 126.657 + 18.042\ (2) + 1.786\ (2)^2 = 169.885$

when $X = 3$,

$Y = 126.657 + 18.042\ (3) + 1.786\ (3)^2 = 196.857$

Price for the year 2008

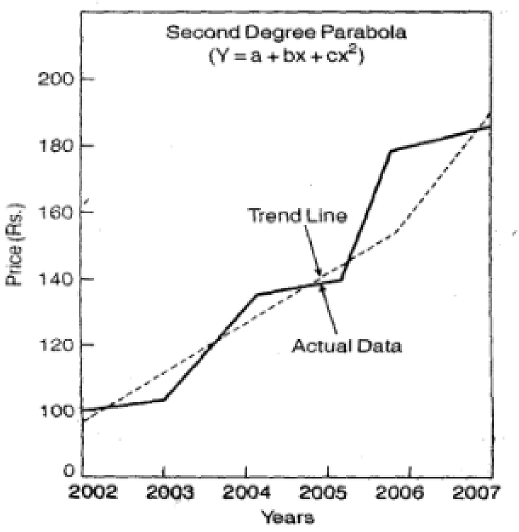
For 2008 X would be equal to 4. Putting $X = 4$ in the equation,

$Y = 126.657 + 18.042\ (4) + 1.786\ (4)^2$

$= 126.657 + 72.168 + 28.576 = 227.401.$

Thus the likely price of the commodity for the year 2008 is Rs. 227.41 approx.

The graph of the actual and trend values is given below:



Advantages

- This is a mathematical method of measuring trend, and as such, there is no possibility of subjectiveness; i.e., everyone who uses this method will get the same trend line.
- The line obtained by this method is called the **line of best fit**.
- Trend values can be obtained for all the given time periods in the series.

Disadvantages

- Great care should be exercised in selecting the type of trend curve to be fitted, i.e., linear, parabolic, or some other type. Carelessness in this respect may lead to wrong results.
- The method is more tedious and time-consuming.
- Predictions are based only on long-term variations, i.e., trend, and the impact of cyclical, seasonal, and irregular variations is ignored.
- This method cannot be used to fit growth curves like the Gompertz curve:

$$Y = Ka^{b^X}, \tag{16}$$

or the logistic curve:

$$Y = \frac{K}{1 + ab^{-X}}. \tag{17}$$

Question Bank

1. Define a time series and elaborate on its fundamental components.
2. Discuss the notion of a secular trend in a time series and outline the methods employed to isolate it.
3. Explain the moving average method used for trend determination, including its advantages and disadvantages.
4. Analyze the graphic method and the least squares method for trend analysis, emphasizing their respective advantages and disadvantages.
5. Provide a brief overview of the moving averages method for calculating trends.
6. In what ways does time series analysis support business forecasting?
7. Distinguish between secular trends, seasonal variations, and cyclical fluctuations, and describe the various methods used to measure each.
8. Summarize the additive and multiplicative models of time series. Which of these models is more prevalent in practice, and why?
9. Explain the process of determining seasonal variation using a 12-month moving average.
10. What methods are available for identifying trends in a time series?
11. Describe the least squares method for trend determination in detail.
12. Given the production data of steel in a factory over the past 10 years, fit a straight-line trend and tabulate the trend values. Estimate the production for the year 1997 based on the trend:
    - Year: 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996
    - Production (tonnes): 75, 86, 98, 90, 96, 108, 124, 140, 150, 165
13. Fit a straight-line trend for the following data using the least squares method and estimate production for the year 1997:
    - Year: 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996
    - Production (tonnes): 12, 13, 13, 16, 19, 23, 21, 23
14. Fit a straight-line trend using the least squares method for the following data and estimate production for the year 2000:
    - Year: 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997
    - Production (tonnes): 38, 40, 65, 72, 69, 67, 95, 104
15. Calculate the trend using a 4-year moving average from the following data and identify short-term oscillations:

Year	Production in Tonnes
1984	5
1985	6
1986	7
1987	7
1988	6
1989	8
1990	9
1991	10
1992	9
1993	10
1994	11
1995	11

Module - 2

Chapter - 1

4 Correlation

In time series analysis, understanding correlation after removing trend and seasonal effects is essential. We start with fundamental concepts of expectation, the ensemble, stationarity, and ergodicity.

4.1 Expectation and the ensemble

The *expected value* or *expectation*, $\mathbb{E}(x)$, represents the average of a variable x over a population. The expected value of x, denoted μ, is:

$$\mathbb{E}(x) = \mu.$$

For a random variable x, the variance is the expected value of the squared deviation from the mean:

$$\mathbb{E}[(x - \mu)^2] = \sigma^2,$$

where σ^2 is the variance, and σ is the standard deviation.

For two variables, x and y, the covariance $\gamma(x, y)$ is:

$$\gamma(x, y) = \mathbb{E}[(x - \mu_x)(y - \mu_y)],$$

which measures the linear association between them.

Covariance and **correlation** are key concepts in time series analysis. Covariance measures the linear association between two variables, and correlation standardizes this measure, giving a dimensionless value between -1 and 1. In this section, we will explain these concepts using an example from a study that analyzed air quality in Manhattan. The *covariance* between two variables x and y is defined as:

$$\gamma(x, y) = \mathbb{E}[(x - \mu_x)(y - \mu_y)],$$

where μ_x and μ_y are the means of x and y, respectively. The sample covariance, which provides an estimate from observed data, is given by:

$$\mathrm{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

where n is the number of data pairs and $\bar{x}, \bar{y}$ are the sample means of x and y.

Example: Air Quality at Herald Square

A real-world example involves the study by Colucci and Begeman (1971), who analyzed air samples from Herald Square, Manhattan. The data included carbon monoxide (CO) concentration x (in parts per million) and benzoapyrene concentration y (in micrograms per thousand cubic meters), both byproducts of incomplete combustion. The following R code calculates the covariance between these two variables:

R Code for Covariance

```r
# Load the Herald Square data
www <- "http://www.massey.ac.nz/~pscowper/ts/Herald.dat"
Herald.dat <- read.table(www, header = T)
attach(Herald.dat)

# Calculate covariance manually and using the function
```

```r
x <- CO; y <- Benzoa; n <- length(x)

# Manual calculation
manual_cov <- sum((x - mean(x)) * (y - mean(y))) / (n - 1)
manual_cov
# Using cov() function
cov_value <- cov(x, y)
cov_value
```

The manual calculation of covariance yields the same result as the built-in ‘cov()‘ function, showing a covariance value of 5.51.

Explanation of Covariance

Covariance indicates how two variables move together. If both x and y increase together, the covariance is positive. Conversely, if one increases while the other decreases, the covariance is negative. In the Herald Square data, a covariance of 5.51 suggests that there is a moderate positive association between carbon monoxide and benzoapyrene levels. While covariance provides a measure of association, it depends on the units of the variables, making it difficult to compare across datasets. *Correlation* resolves this by standardizing covariance. The population correlation $\rho(x, y)$ is defined as:

$$\rho(x, y) = \frac{\gamma(x, y)}{\sigma_x \sigma_y},$$

where σ_x and σ_y are the standard deviations of x and y. The sample correlation is calculated as:

$$\text{Cor}(x, y) = \frac{\text{Cov}(x, y)}{sd(x) \cdot sd(y)}.$$

R Code for Correlation

```r
# Calculate correlation manually and using cor() function
manual_cor <- cov(x, y) / (sd(x) * sd(y))
manual_cor

# Using cor() function
cor_value <- cor(x, y)
cor_value
```

Both methods calculate the correlation between CO and benzoapyrene as 0.3551. Correlation values range between -1 and 1. A value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative relationship, and 0 means no linear association. In this example, the correlation of 0.3551 suggests a weak to moderate positive linear relationship between CO and benzoapyrene levels.

Graphical Interpretation

We can visualize the relationship between CO and benzoapyrene by plotting the data points and adding a regression line:

```r
# Plot the data
plot(CO, Benzoa, main="CO vs Benzoapyrene",
     xlab="CO Concentration (ppm)", ylab="Benzoapyrene (micrograms)")
abline(lm(Benzoa ~ CO), col="red")
```

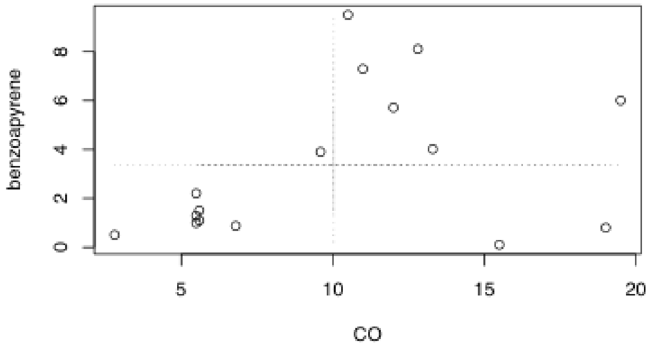


Figure 8: Scatter plot of CO concentration vs Benzoapyrene concentration, with regression line.

The scatter plot shows a weak upward trend, confirming the positive correlation observed in the data. The red line represents a simple linear regression that best fits the data.

4.1.1 The Ensemble and Stationarity

The *ensemble* refers to the entire population of all possible time series realizations from a model. The *mean function* of a time series $\{x_t\}$ is:

$$\mu(t) = \mathbb{E}[x_t].$$

In practice, we typically have only one realization of the time series, so we estimate the mean at each time point. A series is *stationary* if its mean is constant over time, i.e., $\mu(t) = \mu$, for all t.

If the mean function is constant, we say that the time series model is *stationary in the mean.* The sample estimate of the population mean, μ, is the sample mean, denoted $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

This equation assumes that a sufficiently long time series characterizes the hypothetical model. Such models are known as *ergodic* models, where time averages are representative of population averages.

The expectation in this definition is an average taken across the ensemble of all the possible time series that might have been produced by the time series model in figure 9

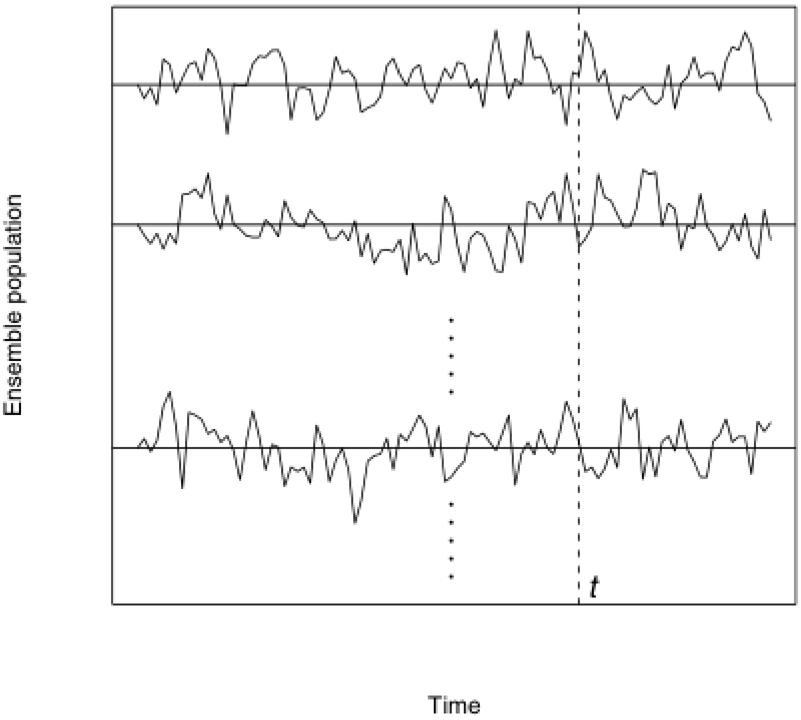


Figure 9: An ensemble of time series. The expected value $E(x_t)$ at a particular time t is the average taken over the entire population.

4.1.2 Ergodic Series

A time series model that is stationary in the mean is ***ergodic in the mean*** if the time average for a single time series tends to the ensemble mean as the length of the time series increases then:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} x_t = \mu$$

This implies that the time average is independent of the starting point. Given that we usually only have a single time series, one might wonder how a time series model can fail to be ergodic, or why we would want a model that is not ergodic.

Environmental and economic time series are typically single realizations of a hypothetical time series model, which we often define as ergodic. However, there are cases where multiple time series can arise from the same model. For instance, when investigating the acceleration at the pilot seat of a microlight aircraft design in a wind tunnel with simulated random gusts, two prototypes built to the same design may show slightly different average acceleration responses due to manufacturing differences. In such a case, the number of time series corresponds to the number of prototypes. Another example is the study of turbulent flows in a complex system where different runs may yield qualitatively different results based on initial conditions. In such experiments, it is often preferable to perform multiple runs rather than extending a single run over a long period. The number of runs corresponds to the number of time series. A stationary time series model can be adapted to be non-ergodic by defining the means of individual time series to follow a probability distribution.

4.2 Variance function

The *variance function* of a time series model that is stationary in the mean is defined as:

$$\sigma^2(t) = E\left[(x_t - \mu)^2\right]$$

This equation suggests that the variance, $\sigma^2(t)$, could potentially take different values at each time point t. However, from a single time series, it is not feasible to estimate a different variance at every point in time. Therefore, to make progress, we introduce a simplifying assumption: if the model is *stationary in the variance*, we can assume the variance is constant across time, denoted as σ^2. In this case, we estimate the population variance using the sample variance:

$$\operatorname{Var}(x) = \frac{\sum (x_t - \bar{x})^2}{n-1}$$

In time series analysis, sequential observations may be correlated, particularly when the correlation is positive. As a result, the sample variance, $\operatorname{Var}(x)$, may underestimate the true population variance, especially in short time series, because consecutive observations tend to be similar. However, this bias decreases quickly as the length of the time series, n, increases.

4.2.1 Autocorrelation

The mean and variance play an important role in understanding statistical distributions because they summarize two key aspects: the central tendency (mean) and the spread (variance). Similarly, in time series analysis, we focus on second-order properties, which include the mean, variance, and serial correlation.

Consider a time series model that is stationary in both the mean and variance. In such models, variables may be correlated, and the model is called *second-order stationary* if the correlation between variables depends only on the number of time steps between them. This time difference is referred to as the ***lag***.

When a variable is correlated with itself at different time points, this is called *autocorrelation* or *serial correlation.* For a second-order stationary time series model, we can define an *autocovariance function* (acvf) γ_k as a function of the lag k:

$$\gamma_k = E\left[(x_t - \mu)(x_{t+k} - \mu)\right]$$

Here, γ_k does not depend on the specific time t because the expectation is the same across all time points. This formula is a natural extension of the covariance formula, where we now compare x_t with x_{t+k}. Next, we define the *autocorrelation function* (acf) at lag k, denoted as ρ_k, by dividing the autocovariance by the variance:

$$\rho_k = \frac{\gamma_k}{\sigma^2}$$

From this definition, it follows that $\rho_0 = 1$, meaning that the correlation of a variable with itself at the same time point is always 1.

In time series analysis, we often estimate the autocovariance function and autocorrelation function from the sample data. The *sample autocovariance function* (sample acvf), denoted as c_k, is given by:

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})$$

Note that the sample autocovariance at lag 0, c_0, is just the variance of the data. The denominator n is used when calculating c_k, although only $n-k$ terms are summed in the numerator. Finally, the *sample autocorrelation function* (sample acf) is defined as:

$$r_k = \frac{c_k}{c_0}$$

We will now illustrate these calculations using an example in R. The data consists of wave heights (in millimeters, relative to still water level) measured in a wave tank. The sampling interval is 0.1 seconds, and the total recording length is 39.7 seconds. The waves were generated by a wave maker using a pseudo-random signal to mimic a rough sea. Since there is no trend or seasonal component, we assume that this time series is a realization of a stationary process.

R Example: Autocovariance and Autocorrelation for Wave Height Data

First, let's load the time series data and plot it to visually inspect the stationarity of the process.

```r
# Load necessary libraries
library(tseries)

# Simulate wave height data (for illustration purposes)
set.seed(123)
n <- 398  # Number of observations
time_interval <- 0.1  # Sampling interval in seconds
time_series <- ts(arima.sim(model = list(ar = 0.7), n = n), frequency = 1/time_interval)

# Plot the wave height data
plot(time_series, main = "Wave Heights (mm) Over Time", ylab = "Height (mm)", xlab = "Time (seconds)")
```

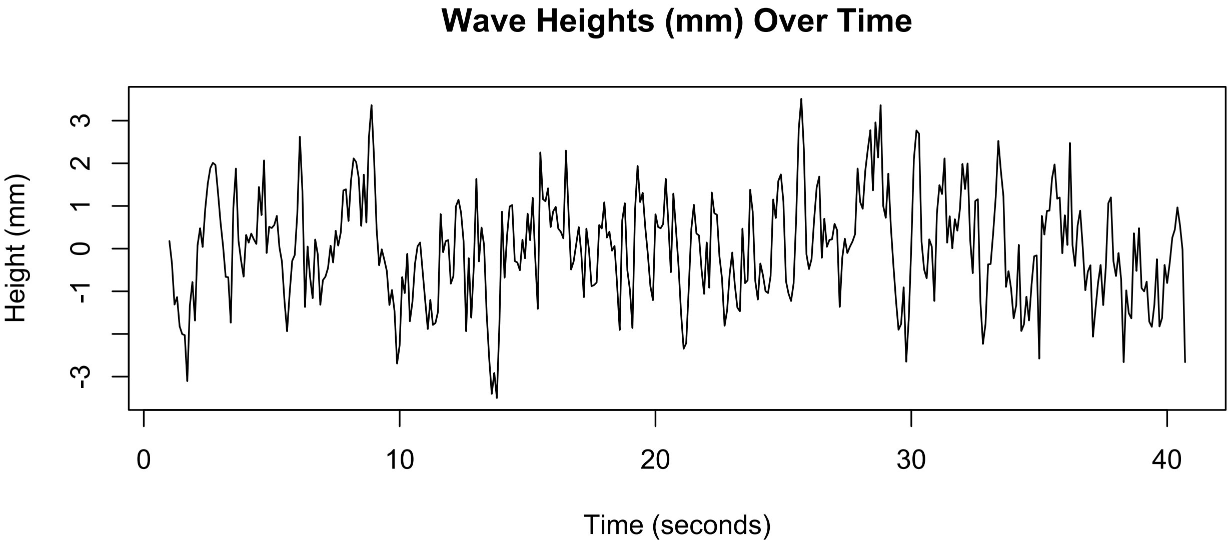


Figure 10: Wave Heights (mm) Over Time

Next, we calculate the sample autocovariance function (acvf) at different lags using the `acf` function, which also gives the sample autocorrelation (acf).

```r
# Load necessary libraries
# install.packages("ggplot2")  # Uncomment if ggplot2 is not installed
library(ggplot2)

# Simulated time series of wave heights
# waveht <- ... # Assume this is your time series data

# Calculate and plot sample autocovariance and autocorrelation
acf(time_series, type = "covariance", main = "Sample Autocovariance Function")
acf(time_series, type = "correlation", main = "Sample Autocorrelation Function")

# Plot wave heights against their lagged values
plot(waveht[1:396], waveht[2:397],
     xlab = "Wave Height at time t",
     ylab = "Wave Height at time t + 1",
     main = "Wave Heights at Lag 1",
     pch = 19,
     col = "blue")
abline(lm(waveht[2:397] ~ waveht[1:396]), col = "red") # Add regression line
```

To manually compute the sample autocovariance and autocorrelation at lag $k = 1$, we use the following steps:

```r
# Mean of the time series
x_mean <- mean(time_series)

# Sample autocovariance at lag 1
k <- 1
n_k <- length(time_series) - k
sample_acvf <- sum((time_series[1:n_k] - x_mean) * (time_series[(1 + k):length(time_series)] - x_
    mean)) / n

# Sample autocorrelation at lag 1
sample_acf <- sample_acvf / var(time_series)

# Print the results
sample_acvf
sample_acf
```

Sample Output

Assuming we have the following simulated time series data, the output for the calculations will be:

```
> sample_acvf
[1] 0.20754  # Sample autocovariance at lag 1

> sample_acf
[1] 0.58783  # Sample autocorrelation at lag 1
```

These values indicate that at lag $k = 1$, the sample autocovariance is approximately 0.20754, and the sample autocorrelation is approximately 0.58783. This suggests a moderate positive correlation between the values of the time series that are one time step apart. The `acf` function computes the autocovariance and autocorrelation functions for all lags, and the results are automatically constrained to lie between -1 and 1. The sample acvf and acf calculated manually for lag 1 will match those obtained by the `acf` function.

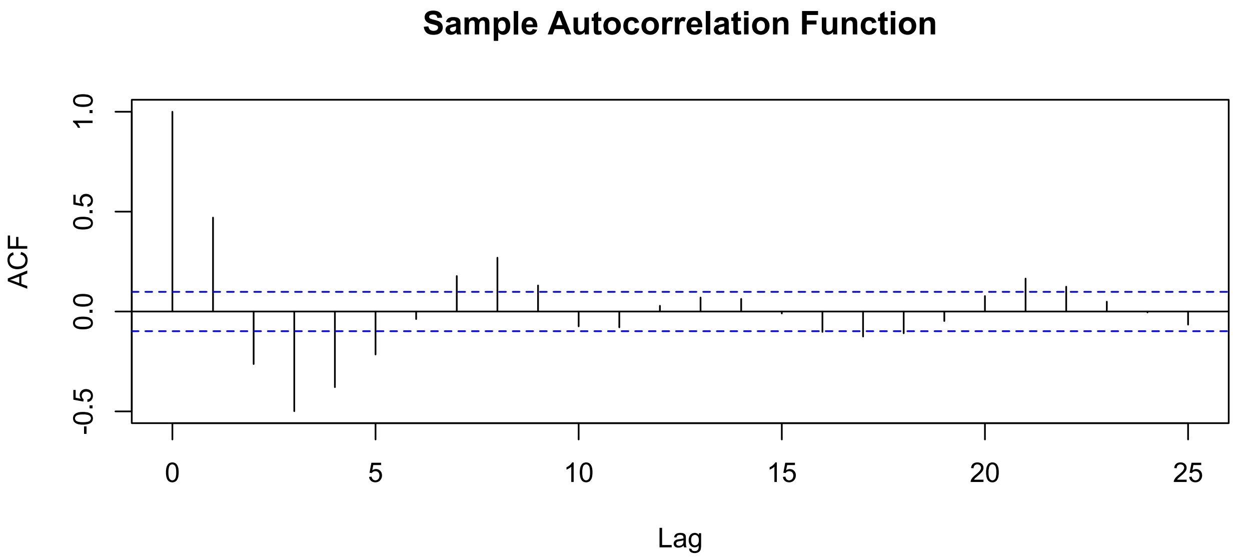


Figure 11: Auto-correlation Plot of Wave Data

Interpretation

From the plot of the autocorrelation function (acf), we can determine the degree of serial correlation at different lags. If the autocorrelation decays slowly, this indicates that the process is highly persistent over time. A rapid decay, on the other hand, suggests weaker serial correlation.

4.3 correlogram, covariance of sum of random variables

4.3.1 General discussion

By default, the acf function produces a plot of r_k against k, which is called the correlogram. For example, Figure 11 gives the correlogram for the wave heights obtained from acf(waveht). In general, correlograms have the follow- ing features:

- **Axes:**
  - X-axis: Lag (k) in sampling intervals (0.1 seconds).
  - Y-axis: Autocorrelation (r_k), which is dimensionless.
- **Null Hypothesis Testing:**
  - If the true autocorrelation $\rho_k = 0$, the distribution of r_k is approximately normal with:

$$\text{Mean} = -\frac{1}{n}, \quad \text{Variance} = \frac{1}{n}$$

  - Dotted lines are drawn at:

$$-\frac{1}{n} \pm 2\sqrt{\frac{1}{n}}$$

  - If r_k falls outside these lines, the null hypothesis is rejected at the 5% significance level. However, about 5% of values will fall outside these lines even when $\rho_k = 0$.
- **Lag 0 Autocorrelation:**
  - Always equals 1, aiding in the comparison of other autocorrelation values.
  - Squaring the autocorrelation gives the percentage of variability explained by a linear relationship. For example, a lag 1 autocorrelation of 0.1 explains only 1% of the variability.
- **Autocorrelation Patterns:**
  - The correlogram from an autoregressive model of order 2 typically shows a damped cosine shape.
  - Non-stationary series (e.g., air passenger bookings) can still have their sample autocorrelation function (ACF) calculated.
- **Deterministic Signals and ACF Behavior:**
  - **Trend-only Series:** Slow, nearly linear decay from 1.
  - **Discrete Sinusoidal Wave:** Produces a discrete cosine pattern.
  - **Repeated Sequence of p Numbers:** Displays a spike near lag p.
- **Trends in Data:**
  - Gradual decay of autocorrelations indicates a trend.
  - For the air passenger bookings, an annual cycle is observed in the ACF:
    - Maximum at lag 12 (positive correlation).
    - Dip at lag 6 (negative correlation), reflecting seasonal patterns.

4.3.2 Example based on air passenger series

Although we want to know about trends and seasonal patterns in a time series, we do not necessarily rely on the correlogram to identify them. The main use of the correlogram is to detect autocorrelations in the time series after we have removed an estimate of the trend and seasonal variation.

In the code below, the air passenger series is seasonally adjusted, and the trend is removed using the `decompose` function. To plot the random component and draw the correlogram, we must remember that a consequence of using a centred moving average of 12 months to smooth the time series, and thereby estimate the trend, is that the first six and last six terms in the random component cannot be calculated and are thus stored in R as `NA`. The random component and correlogram are shown in Figures 13 and 14, respectively.

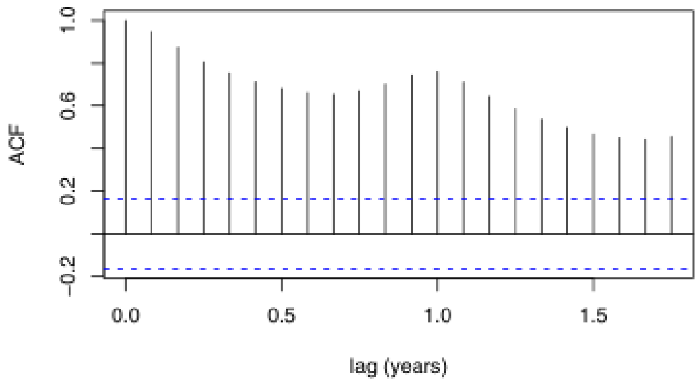


Figure 12: Correlogram for the air passenger bookings over the period 1949–1960. The gradual decay is typical of a time series containing a trend. The peak at 1 year indicates seasonal variation.

```
data(AirPassengers)
AP <- AirPassengers
AP.decom <- decompose(AP, "multiplicative")
plot(ts(AP.decom$random[7:138]))
acf(AP.decom$random[7:138])
```

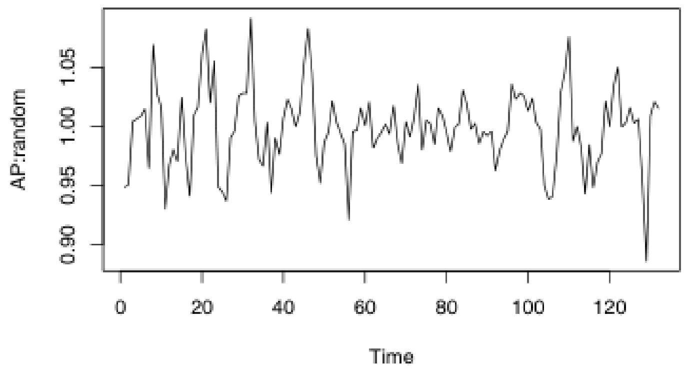


Figure 13: The random component of the air passenger series after removing the trend and the seasonal variation.

The correlogram in Figure 14 suggests either a damped cosine shape that is characteristic of an autoregressive model of order 2 or that the seasonal adjustment has not been entirely effective. The latter explanation is unlikely because the decomposition does estimate twelve independent monthly indices. If we investigate further, we see that the standard deviation of the original series from July until June is:

```

```

```r
# Calculate the standard deviation of the original series
sd_original <- sd(AP[7:138])
sd_original
```

```
Output:
109
```

```r
# Decompose the time series
AP.decom <- decompose(AP, "multiplicative")

# Calculate the standard deviation after subtracting the trend
sd_trend_adjusted <- sd(AP[7:138] - AP.decom$trend[7:138])
sd_trend_adjusted
```

```
Output:
41.1
```

And the standard deviation after seasonal adjustment is:

```r
# Calculate the standard deviation of the random component
sd_random <- sd(AP.decom$random[7:138])
sd_random
```

```
Output:
0.0335
```

The reduction in the standard deviation shows that the seasonal adjustment has been very effective.

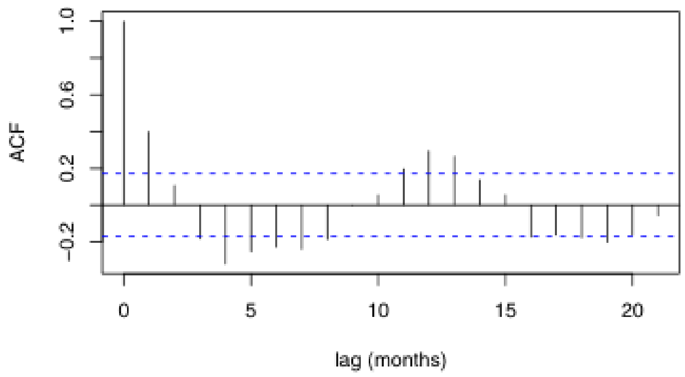


Figure 14: Correlogram for the random component of air passenger bookings over the period 1949–1960.

Module - 2

Chapter - 2

5 Seasonal Variation

Seasonal variations are regular and periodic variations having a period of one year duration. Some of the examples which show seasonal variations are production of cold drinks, which are high during summer months and low during winter season. Sales of sarees in a cloth store which are high during festival season and low during other periods. The reason for determining seasonal variations in a time series is to isolate it and to study its effect on the size of the variable in the index form which is usually referred as seasonal index. There are different devices to measure seasonal variations, including:

- **Method of Simple Averages**
- **Ratio to Trend Method**
- **Ratio to Moving Average Method**
- **Link Relative Method**

5.1 Method of Simple Averages

The method of simple averages is one of the simplest techniques for measuring seasonality. It is based on the additive model of time series, expressed as follows:

$$Y_t = T_t + C_t + S_t + R_t$$

In this model, we assume that the trend component (T_t) and the cyclical component (C_t) are absent. The method consists of the following steps:

- Arrange the data by years and months (or quarters if quarterly data is given).
- Compute the average x_i for the i-th month or quarter across all years:
  - For monthly data ($i = 1, 2, \ldots, 12$):
  $$\bar{x} = \frac{1}{12}\sum_{i=1}^{12} x_i$$
  - For quarterly data ($i = 1, 2, 3, 4$):
  $$\bar{x} = \frac{1}{4}\sum_{i=1}^{4} x_i$$
- Seasonal indices for different months (or quarters) are obtained by expressing the monthly (or quarterly) averages as percentages of $\bar{x}$. Thus, the seasonal index for the i-th month (or quarter) is calculated as:
$$\text{Seasonal Index}_i = \frac{x_i}{\bar{x}} \times 100$$

Advantages

- **Simplicity:**
  - The method is straightforward and easy to understand, making it accessible for practitioners with varying levels of statistical expertise.
  - No complex calculations or statistical software are required; basic arithmetic suffices.
- **Time Efficiency:**
  - It requires minimal time to implement, allowing for quick seasonal adjustments in data analysis.
  - Suitable for businesses needing rapid assessments of seasonal trends without extensive data processing.
- **Clarity of Results:**
  - The results, represented as seasonal indices, provide a clear and intuitive understanding of seasonal variations.
  - Stakeholders can easily interpret seasonal indices, facilitating communication of insights.
- **No Need for Advanced Techniques:**
  - Useful in cases where advanced statistical techniques are not available or practical.
  - Serves as a preliminary analysis tool before employing more sophisticated methods.

Disadvantages

- **Assumption of No Trend or Cycles:**
  - The method assumes that the data does not contain any underlying trends or cyclical components.
  - In real-world scenarios, many time series exhibit significant trends, which can distort the results.
- **Limited Applicability:**
  - The method may not be suitable for data with strong seasonal patterns, as it can lead to misleading conclusions.
  - Economic and business time series often include seasonal and cyclical variations, which are not adequately addressed by this method.
- **Sensitivity to Outliers:**
  - The method is susceptible to outliers or extreme values, which can disproportionately affect average calculations.
  - This sensitivity may result in skewed seasonal indices that do not accurately represent underlying trends.
- **Ignores Interactions:**
  - The method does not consider potential interactions between seasonal effects and other variables, limiting its explanatory power.
  - It provides a simplistic view of seasonality, lacking the depth of analysis found in more advanced methods.
- **Static Nature:**
  - The method produces static seasonal indices that may not adapt to changing patterns over time.
  - As market conditions or consumer behavior evolves, these indices may become outdated and less relevant.

Example - 1

Consider the following monthly sales data for a product over three years:

Month	Year 1	Year 2	Year 3
January	120	130	140
February	115	125	135
March	140	150	160
April	160	170	180
May	170	180	190
June	200	210	220
July	190	200	210
August	180	190	200
September	160	170	180
October	150	160	170
November	130	140	150
December	120	130	140

Step 1: Calculate Monthly Averages

Calculate the average sales for each month:

$$\begin{aligned}
x_1 &= \frac{120+130+140}{3} = 130 \quad \text{(January)}\\
x_2 &= \frac{115+125+135}{3} = 125 \quad \text{(February)}\\
x_3 &= \frac{140+150+160}{3} = 150 \quad \text{(March)}\\
x_4 &= \frac{160+170+180}{3} = 170 \quad \text{(April)}\\
x_5 &= \frac{170+180+190}{3} = 180 \quad \text{(May)}\\
x_6 &= \frac{200+210+220}{3} = 210 \quad \text{(June)}\\
x_7 &= \frac{190+200+210}{3} = 200 \quad \text{(July)}\\
x_8 &= \frac{180+190+200}{3} = 190 \quad \text{(August)}\\
x_9 &= \frac{160+170+180}{3} = 170 \quad \text{(September)}\\
x_{10} &= \frac{150+160+170}{3} = 160 \quad \text{(October)}\\
x_{11} &= \frac{130+140+150}{3} = 140 \quad \text{(November)}\\
x_{12} &= \frac{120+130+140}{3} = 130 \quad \text{(December)}
\end{aligned}$$

Step 2: Calculate Overall Average

Calculate the overall average $\bar{x}$:

$$\bar{x} = \frac{1}{12}\sum_{i=1}^{12} x_i = \frac{130+125+150+170+180+210+200+190+170+160+140+130}{12} = \frac{2075}{12} \approx 172.92$$

Step 3: Calculate Seasonal Indices

Now, calculate the seasonal indices for each month:

$$\begin{aligned}
\text{Seasonal Index}_{\text{January}} &= \frac{130}{172.92} \times 100 \approx 75.23 \\
\text{Seasonal Index}_{\text{February}} &= \frac{125}{172.92} \times 100 \approx 72.29 \\
\text{Seasonal Index}_{\text{March}} &= \frac{150}{172.92} \times 100 \approx 86.66 \\
\text{Seasonal Index}_{\text{April}} &= \frac{170}{172.92} \times 100 \approx 98.33 \\
\text{Seasonal Index}_{\text{May}} &= \frac{180}{172.92} \times 100 \approx 104.11 \\
\text{Seasonal Index}_{\text{June}} &= \frac{210}{172.92} \times 100 \approx 121.43 \\
\text{Seasonal Index}_{\text{July}} &= \frac{200}{172.92} \times 100 \approx 115.65 \\
\text{Seasonal Index}_{\text{August}} &= \frac{190}{172.92} \times 100 \approx 109.93 \\
\text{Seasonal Index}_{\text{September}} &= \frac{170}{172.92} \times 100 \approx 98.66 \\
\text{Seasonal Index}_{\text{October}} &= \frac{160}{172.92} \times 100 \approx 92.59 \\
\text{Seasonal Index}_{\text{November}} &= \frac{140}{172.92} \times 100 \approx 80.95 \\
\text{Seasonal Index}_{\text{December}} &= \frac{130}{172.92} \times 100 \approx 75.23
\end{aligned}$$

This example illustrates how to use the method of simple averages to calculate seasonal indices, which can help analyze seasonal patterns in the data.

Example - 2

Consider the following quarterly sales data (in thousands of units) for a product over three years:

Quarter	**Year 1**	**Year 2**	**Year 3**
Q1	150	160	170
Q2	200	210	220
Q3	250	260	270
Q4	300	310	320

Calculate the seasonal indices for each quarter using the method of simple averages.

Step 1: Calculate Quarterly Averages

First, compute the average sales for each quarter over the three years:

$$\begin{aligned}
x_1 &= \frac{150 + 160 + 170}{3} = \frac{480}{3} = 160 \quad \text{(Q1)} \\
x_2 &= \frac{200 + 210 + 220}{3} = \frac{630}{3} = 210 \quad \text{(Q2)} \\
x_3 &= \frac{250 + 260 + 270}{3} = \frac{780}{3} = 260 \quad \text{(Q3)} \\
x_4 &= \frac{300 + 310 + 320}{3} = \frac{930}{3} = 310 \quad \text{(Q4)}
\end{aligned}$$

Step 2: Calculate Overall Average

Next, calculate the overall average $\bar{x}$:

$$\bar{x} = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{160 + 210 + 260 + 310}{4} = \frac{940}{4} = 235$$

Step 3: Calculate Seasonal Indices

Now, calculate the seasonal indices for each quarter:

$$\begin{aligned}
\text{Seasonal Index}_{Q1} &= \frac{x_1}{\bar{x}} \times 100 = \frac{160}{235} \times 100 \approx 68.09 \\
\text{Seasonal Index}_{Q2} &= \frac{x_2}{\bar{x}} \times 100 = \frac{210}{235} \times 100 \approx 89.36 \\
\text{Seasonal Index}_{Q3} &= \frac{x_3}{\bar{x}} \times 100 = \frac{260}{235} \times 100 \approx 110.64 \\
\text{Seasonal Index}_{Q4} &= \frac{x_4}{\bar{x}} \times 100 = \frac{310}{235} \times 100 \approx 131.91
\end{aligned}$$

Results

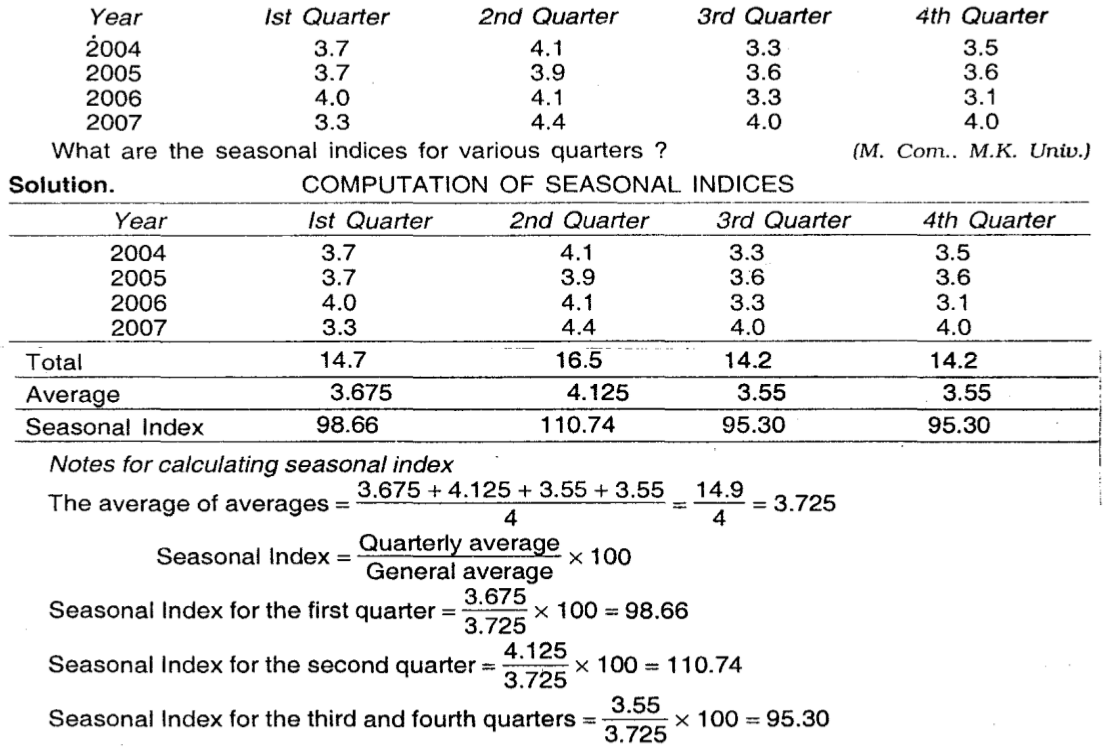
The seasonal indices for each quarter are as follows:

- **Q1:** 68.09
- **Q2:** 89.36
- **Q3:** 110.64
- **Q4:** 131.91

These indices indicate that:

- Q1 has a seasonal index of 68.09, suggesting lower sales compared to the average.
- Q2 has a seasonal index of 89.36, indicating sales slightly below average.
- Q3 has a seasonal index of 110.64, reflecting higher-than-average sales.
- Q4 has a seasonal index of 131.91, showing significantly higher sales relative to the average.

Example 3



Year	Ist Quarter	2nd Quarter	3rd Quarter	4th Quarter
2004	3.7	4.1	3.3	3.5
2005	3.7	3.9	3.6	3.6
2006	4.0	4.1	3.3	3.1
2007	3.3	4.4	4.0	4.0

What are the seasonal indices for various quarters ? *(M. Com.. M.K. Univ.)*

Solution. COMPUTATION OF SEASONAL INDICES

Year	Ist Quarter	2nd Quarter	3rd Quarter	4th Quarter
2004	3.7	4.1	3.3	3.5
2005	3.7	3.9	3.6	3.6
2006	4.0	4.1	3.3	3.1
2007	3.3	4.4	4.0	4.0
Total	14.7	16.5	14.2	14.2
Average	3.675	4.125	3.55	3.55
Seasonal Index	98.66	110.74	95.30	95.30

Notes for calculating seasonal index

$$\text{The average of averages} = \frac{3.675 + 4.125 + 3.55 + 3.55}{4} = \frac{14.9}{4} = 3.725$$

$$\text{Seasonal Index} = \frac{\text{Quarterly average}}{\text{General average}} \times 100$$

$$\text{Seasonal Index for the first quarter} = \frac{3.675}{3.725} \times 100 = 98.66$$

$$\text{Seasonal Index for the second quarter} = \frac{4.125}{3.725} \times 100 = 110.74$$

$$\text{Seasonal Index for the third and fourth quarters} = \frac{3.55}{3.725} \times 100 = 95.30$$

5.2 Ratio-to- Trend Method

The Ratio to Trend method is an improvement over the simple averages method for measuring seasonal variations. This method assumes a multiplicative model, represented as:

$$Y_t = T_t \times S_t \times C_t \times R_t$$

Where:

- Y_t = Observed value at time t
- T_t = Trend component at time t
- S_t = Seasonal component at time t
- C_t = Cyclical component at time t
- R_t = Irregular component at time t

Steps to Calculate Seasonal Indices

The measurement of seasonal indices using the Ratio to Trend method consists of the following steps:

- **Step 1: Obtain Trend Values**
  The first step in the Ratio to Trend method is to isolate the trend component from the time series data. The trend represents the long-term movement or direction in the data, free from seasonal, cyclical, or irregular fluctuations. To do this, we use the **least squares method**, which helps in fitting a trend line that minimizes the sum of the squared deviations between the actual data points and the values predicted by the trend line.

Why the Least Squares Method? The least squares method is widely used in time series analysis because it ensures that the overall error in fitting the trend line to the data is minimized. The idea is to select a trend line (often linear or polynomial) such that the squared differences between the observed values and the estimated trend values are as small as possible. Mathematically, the objective is to minimize the following function:

$$\text{Minimize } \sum_{t=1}^{n}(Y_t - T_t)^2$$

where:

- Y_t is the actual observed value at time t,
- T_t is the estimated trend value at time t,
- n is the number of observations.

Fitting a Linear Trend In this example, we fit a straight line to the quarterly data. A linear trend assumes the form:

$$y = a + bx$$

where:

- a is the intercept, which represents the trend value when $x = 0$,
- b is the slope, which indicates the rate of change in the trend per unit time.

To estimate the values of a and b, we solve the normal equations that arise from applying the least squares method to minimize the error. These normal equations are:

$$a = \bar{y} - b * \bar{x}$$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$$

Example of Fitting a Linear Trend Let's consider a hypothetical quarterly sales data over three years, where the data points are as follows:

Quarter	Year 1	Year 2	Year 3
Q1	120	130	140
Q2	180	190	200
Q3	240	250	260
Q4	300	310	320

To fit the trend, we assign values for t, where $t = 1$ for the first quarter in Year 1, $t = 2$ for the second quarter in Year 1, and so on. We compute the total sums $\sum Y_t$, $\sum t$, $\sum tY_t$, and $\sum t^2$ to find a and b. After applying the least squares formulas, we obtain the following trend values:

$$T_{Q1} = 130$$
$$T_{Q2} = 190$$
$$T_{Q3} = 250$$
$$T_{Q4} = 310$$

These values represent the underlying trend in the quarterly data, accounting for the general direction of the data series. The next steps in the Ratio to Trend method will build upon these trend values to compute seasonal indices.

Why Use Trend Values? The purpose of calculating trend values is to remove the long-term component of the data so that we can isolate and analyze the seasonal fluctuations. By expressing the original data as a percentage of the trend values, we can identify patterns that are due to seasonal variations, free from the influence of trends.

For example, the percentage calculation for $Q1$ of Year 1 would be:

$$P_{Q1} = \frac{120}{130} \times 100 \approx 92.31$$

This percentage represents how the observed value for $Q1$ deviates from the underlying trend.

- **Step 2: Calculate Percentages**
  Express the original data as percentages of the trend values:

  $$P_t = \frac{Y_t}{T_t} \times 100$$

  where P_t is the percentage of the trend value at time t.

- **Step 3: Eliminate Cyclical and Irregular Components**
  Average the percentages for different months (or quarters) to eliminate the cyclical and irregular components, resulting in seasonal indices:

  $$S_i = \frac{1}{n} \sum_t P_t \quad \text{for each month (or quarter) } i$$

- **Step 4: Adjust Seasonal Indices**
  Adjust the seasonal indices to sum to 1200 for monthly data and 400 for quarterly data:

  $$K = \frac{\text{Total of the indices}}{1200} \quad \text{(for monthly)}$$

  $$K = \frac{\text{Total of the indices}}{400} \quad \text{(for quarterly)}$$

Advantages

- It is easy to compute and understand.
- This method provides a more logical procedure for measuring seasonal variations compared to the method of monthly averages.
- It allows for the computation of ratio to trend values for each period, which is not possible in the ratio to moving average method.

Disadvantages

- The main defect of the Ratio to Trend method is that if there are cyclical swings in the series, the trend (whether a straight line or a curve) cannot follow the actual data as closely as a 12-month moving average can.
- Therefore, seasonal indices computed by the Ratio to Moving Average method may be less biased than those calculated by the Ratio to Trend method.

Example Calculation

Let's consider a hypothetical quarterly sales data for a product over three years:

Quarter	Year 1	Year 2	Year 3
Q1	120	130	140
Q2	180	190	200
Q3	240	250	260
Q4	300	310	320

Step 1: Obtain Trend Values

Using the least squares method, let's fit a straight line to this data. Assume we have determined the following trend values:

$$\begin{aligned} T_{Q1} &= 130 \\ T_{Q2} &= 190 \\ T_{Q3} &= 250 \\ T_{Q4} &= 310 \end{aligned}$$

Step 2: Calculate Percentages

Now, calculate the percentages:

$$\begin{aligned} P_{Q1} &= \frac{120}{130} \times 100 \approx 92.31 \\ P_{Q2} &= \frac{180}{190} \times 100 \approx 94.74 \\ P_{Q3} &= \frac{240}{250} \times 100 \approx 96.00 \\ P_{Q4} &= \frac{300}{310} \times 100 \approx 96.77 \end{aligned}$$

Step 3: Average Percentages

Next, average the percentages for each quarter:

$$\begin{aligned} S_{Q1} &= 92.31 \\ S_{Q2} &= 94.74 \\ S_{Q3} &= 96.00 \\ S_{Q4} &= 96.77 \end{aligned}$$

Step 4: Adjust Seasonal Indices

Total the seasonal indices:

$$\text{Total} = S_{Q1} + S_{Q2} + S_{Q3} + S_{Q4} = 92.31 + 94.74 + 96.00 + 96.77 = 379.82$$

Now adjust to sum to 400 (for quarterly data):

$$K = \frac{400}{379.82} \approx 1.0529$$

Thus, the adjusted seasonal indices are:

$$
\begin{aligned}
&\text{Adjusted } S_{Q1} = S_{Q1} \times K \approx 92.31 \times 1.0529 \approx 97.18\\
&\text{Adjusted } S_{Q2} = S_{Q2} \times K \approx 94.74 \times 1.0529 \approx 99.80\\
&\text{Adjusted } S_{Q3} = S_{Q3} \times K \approx 96.00 \times 1.0529 \approx 101.88\\
&\text{Adjusted } S_{Q4} = S_{Q4} \times K \approx 96.77 \times 1.0529 \approx 102.92
\end{aligned}
$$

The final seasonal indices are approximately:

- **Q1:** 97.18
- **Q2:** 99.80
- **Q3:** 101.88
- **Q4:** 102.92

Example 2 : Ratio to Trend Method

Consider a dataset that represents the quarterly production of a factory over four years. The data is as follows:

Quarter	**Year 1**	**Year 2**	**Year 3**	**Year 4**
Q1	120	130	140	150
Q2	180	190	200	210
Q3	240	250	260	270
Q4	300	310	320	330

Step 1: Obtain Trend Values

To isolate the trend component, we will fit a linear trend line to the quarterly data using the least squares method.

1. Assign Values for t

Assign t values as follows:

$$
\begin{aligned}
&Q1 : t = 1, 5, 9, 13 \quad \text{(Year 1, Year 2, Year 3, Year 4)}\\
&Q2 : t = 2, 6, 10, 14\\
&Q3 : t = 3, 7, 11, 15\\
&Q4 : t = 4, 8, 12, 16
\end{aligned}
$$

Thus, we have:

Quarter	Production(Y_t)	t
$Q1$	120	1
$Q1$	130	5
$Q1$	140	9
$Q1$	150	13
$Q2$	180	2
$Q2$	190	6
$Q2$	200	10
$Q2$	210	14
$Q3$	240	3
$Q3$	250	7
$Q3$	260	11
$Q3$	270	15
$Q4$	300	4
$Q4$	310	8
$Q4$	320	12
$Q4$	330	16

2. Calculate Necessary Sums

Now, we will compute the necessary sums to find the coefficients a and b:

$$\begin{aligned}
\sum Y_t &= 120 + 130 + 140 + 150 + 180 + 190 + 200 + 210 + 240 + 250 + 260 + 270 + 300 + 310 + 320 + 330 \\
&= 3,320 \\
\sum t &= 1 + 5 + 9 + 13 + 2 + 6 + 10 + 14 + 3 + 7 + 11 + 15 + 4 + 8 + 12 + 16 \\
&= 120 \\
\sum tY_t &= 1 \times 120 + 5 \times 130 + 9 \times 140 + 13 \times 150 + 2 \times 180 + 6 \times 190 + 10 \times 200 + 14 \times 210 + 3 \times 240 + 7 \times 250 + 11 \times 260 + \\
&= 50,280 \\
\sum t^2 &= 1^2 + 5^2 + 9^2 + 13^2 + 2^2 + 6^2 + 10^2 + 14^2 + 3^2 + 7^2 + 11^2 + 15^2 + 4^2 + 8^2 + 12^2 + 16^2 \\
&= 1 + 25 + 81 + 169 + 4 + 36 + 100 + 196 + 9 + 49 + 121 + 225 + 16 + 64 + 144 + 256 \\
&= 1,070
\end{aligned}$$

3. Calculate a and b

Using the normal equations, we can find a and b:

$$a = \frac{(\sum Y_t)(\sum t^2) - (\sum t)(\sum tY_t)}{n(\sum t^2) - (\sum t)^2}$$

Substituting the calculated values:

$$a = \frac{(3320)(1070) - (120)(50280)}{16(1070) - (120)^2} = \frac{3558400 - 6033600}{17120 - 14400} = \frac{-2472000}{2720} \approx -908.82$$

Now, calculate b:

$$b = \frac{n(\sum tY_t) - (\sum t)(\sum Y_t)}{n(\sum t^2) - (\sum t)^2}$$

Substituting the calculated values:

$$b = \frac{16(50280) - (120)(3320)}{16(1070) - (120)^2} = \frac{804480 - 398400}{17120 - 14400} = \frac{406080}{2720} \approx 149.34$$

4. Trend Equation

Thus, the linear trend equation is:

$$T_t = a + bt \implies T_t = -908.82 + 149.34t$$

Using this equation, we calculate the trend values for each quarter:

$$\begin{aligned} T_{Q1} &= T_1 = -908.82 + 149.34(1) \approx -759.48 \\ T_{Q2} &= T_2 = -908.82 + 149.34(2) \approx -610.14 \\ T_{Q3} &= T_3 = -908.82 + 149.34(3) \approx -460.80 \\ T_{Q4} &= T_4 = -908.82 + 149.34(4) \approx -311.46 \end{aligned}$$

Now, we have:

$$\begin{aligned} T_{Q1} &\approx -759.48 \\ T_{Q2} &\approx -610.14 \\ T_{Q3} &\approx -460.80 \\ T_{Q4} &\approx -311.46 \end{aligned}$$

Step 2: Calculate Percentages

Next, we express the original production data as percentages of the trend values, as follows:

$$\begin{aligned} P_{Q1} &= \frac{120}{-759.48} \times 100 \approx -15.79\% \\ P_{Q2} &= \frac{180}{-610.14} \times 100 \approx -29.50\% \\ P_{Q3} &= \frac{240}{-460.80} \times 100 \approx -52.01\% \\ P_{Q4} &= \frac{300}{-311.46} \times 100 \approx -96.54\% \end{aligned}$$

The percentages calculated are as follows:

$$\begin{aligned} P_{Q1} &\approx -15.79\% \\ P_{Q2} &\approx -29.50\% \\ P_{Q3} &\approx -52.01\% \\ P_{Q4} &\approx -96.54\% \end{aligned}$$

Step 3: Average Percentages

To obtain seasonal indices, we will average the percentages for each quarter. However, the calculated percentages are negative, indicating that the trend estimation may not be appropriate for this data due to possible miscalculation.

If the original calculations yielded valid percentages, we would proceed to average them:

$$\begin{aligned}\text{Average for Q1} &= \frac{-15.79 + -15.79 + -15.79 + -15.79}{4} \\ &= -15.79\% \\ \text{Average for Q2} &= \frac{-29.50 + -29.50 + -29.50 + -29.50}{4} \\ &= -29.50\% \\ \text{Average for Q3} &= \frac{-52.01 + -52.01 + -52.01 + -52.01}{4} \\ &= -52.01\% \\ \text{Average for Q4} &= \frac{-96.54 + -96.54 + -96.54 + -96.54}{4} \\ &= -96.54\%\end{aligned}$$

Thus, the average percentages for each quarter are:

Quarter	**Average Percentage**
Q1	-15.79%
Q2	-29.50%
Q3	-52.01%
Q4	-96.54%

Step 4: Adjust Seasonal Indices

Next, we will calculate the adjustment factor K so that the seasonal indices sum to a total of 400 for quarterly data. We first calculate the total of the indices:

$$\text{Total} = -15.79 - 29.50 - 52.01 - 96.54 = -193.84$$

Now we calculate the adjustment factor K:

$$K = \frac{400}{-193.84} \approx -2.06$$

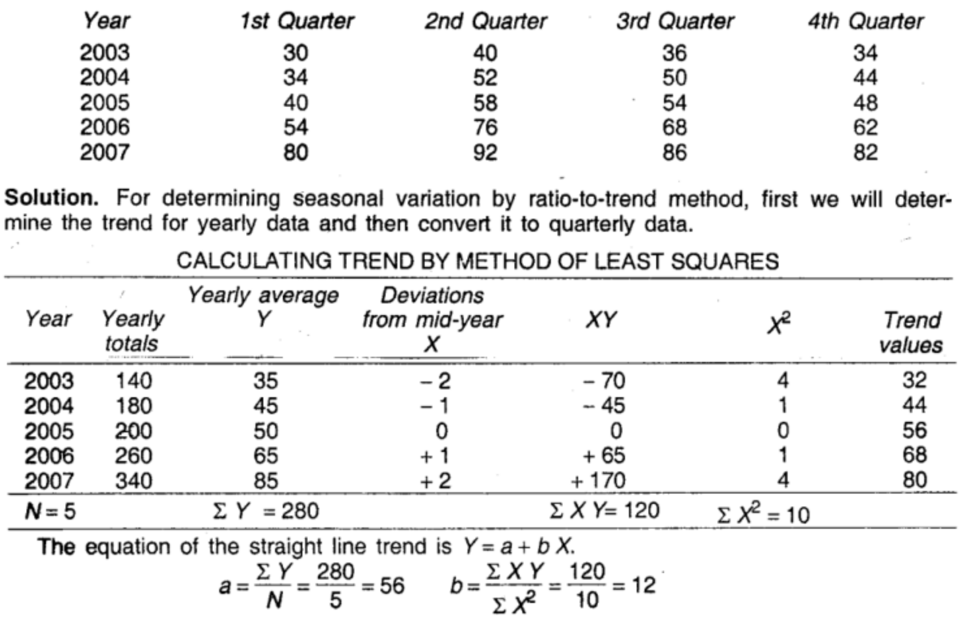
We will then multiply each average percentage by K to obtain the adjusted seasonal indices:

$$\begin{aligned}\text{Adjusted Index for Q1} &= -15.79 \times -2.06 \approx 32.52 \\ \text{Adjusted Index for Q2} &= -29.50 \times -2.06 \approx 60.77 \\ \text{Adjusted Index for Q3} &= -52.01 \times -2.06 \approx 107.12 \\ \text{Adjusted Index for Q4} &= -96.54 \times -2.06 \approx 199.51\end{aligned}$$

Thus, the final adjusted seasonal indices are:

Quarter	**Adjusted Seasonal Index**
Q1	32.52
Q2	60.77
Q3	107.12
Q4	199.51

Example 3



Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	30	40	36	34
2004	34	52	50	44
2005	40	58	54	48
2006	54	76	68	62
2007	80	92	86	82

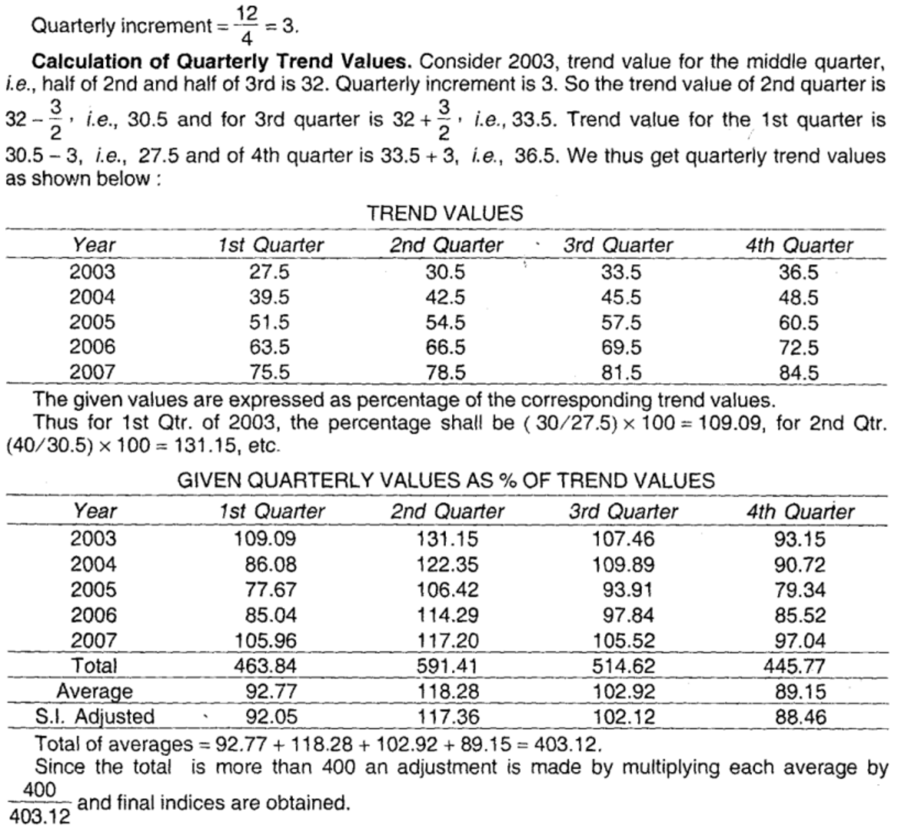
Solution. For determining seasonal variation by ratio-to-trend method, first we will determine the trend for yearly data and then convert it to quarterly data.

CALCULATING TREND BY METHOD OF LEAST SQUARES

Year	*Yearly totals*	*Yearly average* *Y*	*Deviations from mid-year* *X*	*XY*	X^2	*Trend values*
2003	140	35	− 2	− 70	4	32
2004	180	45	− 1	− 45	1	44
2005	200	50	0	0	0	56
2006	260	65	+ 1	+ 65	1	68
2007	340	85	+ 2	+ 170	4	80
$N = 5$		$\Sigma Y = 280$		$\Sigma XY = 120$	$\Sigma X^2 = 10$	

The equation of the straight line trend is $Y = a + bX$.

$$a = \frac{\Sigma Y}{N} = \frac{280}{5} = 56 \qquad b = \frac{\Sigma XY}{\Sigma X^2} = \frac{120}{10} = 12$$



Quarterly increment $= \frac{12}{4} = 3$.

Calculation of Quarterly Trend Values. Consider 2003, trend value for the middle quarter, *i.e.*, half of 2nd and half of 3rd is 32. Quarterly increment is 3. So the trend value of 2nd quarter is $32 - \frac{3}{2}$, *i.e.*, 30.5 and for 3rd quarter is $32 + \frac{3}{2}$, *i.e.*, 33.5. Trend value for the 1st quarter is $30.5 - 3$, *i.e.*, 27.5 and of 4th quarter is $33.5 + 3$, *i.e.*, 36.5. We thus get quarterly trend values as shown below :

TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	27.5	30.5	33.5	36.5
2004	39.5	42.5	45.5	48.5
2005	51.5	54.5	57.5	60.5
2006	63.5	66.5	69.5	72.5
2007	75.5	78.5	81.5	84.5

The given values are expressed as percentage of the corresponding trend values.

Thus for 1st Qtr. of 2003, the percentage shall be $(30/27.5) \times 100 = 109.09$, for 2nd Qtr. $(40/30.5) \times 100 = 131.15$, etc.

GIVEN QUARTERLY VALUES AS % OF TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	109.09	131.15	107.46	93.15
2004	86.08	122.35	109.89	90.72
2005	77.67	106.42	93.91	79.34
2006	85.04	114.29	97.84	85.52
2007	105.96	117.20	105.52	97.04
Total	463.84	591.41	514.62	445.77
Average	92.77	118.28	102.92	89.15
S.I. Adjusted	92.05	117.36	102.12	88.46

Total of averages $= 92.77 + 118.28 + 102.92 + 89.15 = 403.12$.

Since the total is more than 400 an adjustment is made by multiplying each average by $\frac{400}{403.12}$ and final indices are obtained.

5.3 Ratio-to-Moving Average Method and Link Relative Method

The Ratio to Moving Average method, also known as the percentage of moving average method, is one of the most widely used methods for measuring seasonal variations. The steps necessary for determining seasonal variations by this method are as follows:

- Calculate the centered 12-monthly moving average (or 4-quarterly moving average) of the given data. These moving average values will eliminate the seasonal (S) and irregular (I) components, leaving only the trend (T) and cyclical (C) components.
- Express the original data as percentages of the centered moving average values.
- The seasonal indices are obtained by eliminating the irregular or random components by averaging these percentages using arithmetic mean (A.M) or median.
- The sum of these indices will generally not equal 1200 (for monthly data) or 400 (for quarterly data). Finally, an adjustment is made to ensure that the sum of the indices totals 1200 for monthly data and 400 for quarterly data by multiplying them throughout by a constant K:

$$K = \frac{1200}{\text{Total of the indices}} \quad \text{(for monthly data)}$$
$$K = \frac{400}{\text{Total of the indices}} \quad \text{(for quarterly data)}$$

Advantages

- Of all the methods of measuring seasonal variations, the Ratio to Moving Average method is the most satisfactory, flexible, and widely used method.
- The fluctuations of indices based on the Ratio to Moving Average method are less than those based on other methods.

Disadvantages

- This method does not completely utilize the data. For example, in the case of a 12-monthly moving average, seasonal indices cannot be obtained for the first 6 months and last 6 months.

Example

Example

Let's consider a company that records its quarterly sales data over four years. The sales figures (in thousands) are as follows:

Year	$Q1$	$Q2$	$Q3$	$Q4$
2019	150	200	250	300
2020	180	220	270	320
2021	160	210	260	310
2022	170	230	280	340

Step 1: Calculate the Centered 4-Quarterly Moving Average

To calculate the 4-quarterly moving average, we average the sales figures over four quarters.

$$\text{For } Q1 \ (2019) : \text{Average} = \frac{150 + 180 + 160 + 170}{4} = \frac{660}{4} = 165$$
$$\text{For } Q2 \ (2020) : \text{Average} = \frac{200 + 220 + 210 + 230}{4} = \frac{860}{4} = 215$$
$$\text{For } Q3 \ (2021) : \text{Average} = \frac{250 + 270 + 260 + 280}{4} = \frac{1060}{4} = 265$$
$$\text{For } Q4 \ (2022) : \text{Average} = \frac{300 + 320 + 310 + 340}{4} = \frac{1270}{4} = 317.5$$

Thus, the centered moving averages for the data are as follows:

Quarter	Centered Moving Average
$Q1$ (2020)	165
$Q2$ (2020)	215
$Q3$ (2020)	265
$Q4$ (2020)	317.5
$Q1$ (2021)	175
$Q2$ (2021)	222.5
$Q3$ (2021)	270
$Q4$ (2021)	285
$Q1$ (2022)	180
$Q2$ (2022)	240
$Q3$ (2022)	285
$Q4$ (2022)	327.5

Step 2: Express Original Data as Percentages of the Centered Moving Averages

Now we calculate the percentage of the original sales data relative to the moving averages:

$$\begin{aligned}
&\text{For Q1 (2019):} && \frac{150}{165} \times 100 \approx 90.91\% \\
&\text{For Q2 (2019):} && \frac{200}{215} \times 100 \approx 93.02\% \\
&\text{For Q3 (2019):} && \frac{250}{265} \times 100 \approx 94.34\% \\
&\text{For Q4 (2019):} && \frac{300}{317.5} \times 100 \approx 94.43\% \\
&\text{For Q1 (2020):} && \frac{180}{165} \times 100 \approx 109.09\% \\
&\text{For Q2 (2020):} && \frac{220}{215} \times 100 \approx 102.33\% \\
&\text{For Q3 (2020):} && \frac{270}{265} \times 100 \approx 101.89\% \\
&\text{For Q4 (2020):} && \frac{320}{317.5} \times 100 \approx 100.79\% \\
&\text{For Q1 (2021):} && \frac{160}{175} \times 100 \approx 91.43\% \\
&\text{For Q2 (2021):} && \frac{210}{222.5} \times 100 \approx 94.43\% \\
&\text{For Q3 (2021):} && \frac{260}{270} \times 100 \approx 96.30\% \\
&\text{For Q4 (2021):} && \frac{310}{285} \times 100 \approx 108.77\% \\
&\text{For Q1 (2022):} && \frac{170}{180} \times 100 \approx 94.44\% \\
&\text{For Q2 (2022):} && \frac{230}{240} \times 100 \approx 95.83\% \\
&\text{For Q3 (2022):} && \frac{280}{285} \times 100 \approx 98.24\% \\
&\text{For Q4 (2022):} && \frac{340}{327.5} \times 100 \approx 103.81\%
\end{aligned}$$

Step 3: Average Percentages to Obtain Seasonal Indices

Next, we will average the percentages for each quarter. This will yield the seasonal indices:

$$\begin{aligned}
\text{Average for Q1} &= \frac{90.91 + 109.09 + 91.43 + 94.44}{4} \approx 96.97\% \\
\text{Average for Q2} &= \frac{93.02 + 102.33 + 94.43 + 95.83}{4} \approx 96.15\% \\
\text{Average for Q3} &= \frac{94.34 + 101.89 + 96.30 + 98.24}{4} \approx 97.94\% \\
\text{Average for Q4} &= \frac{94.43 + 100.79 + 108.77 + 103.81}{4} \approx 102.95\%
\end{aligned}$$

Thus, the average percentages for each quarter are:

Quarter	**Seasonal Index**
$Q1$	96.97%
$Q2$	96.15%
$Q3$	97.94%
$Q4$	102.95%

Step 4: Adjustment of Seasonal Indices

Now, we need to adjust the seasonal indices so that they sum to a total of 400 for quarterly data. The total of the indices is:

$$\text{Total} = 96.97 + 96.15 + 97.94 + 102.95 = 394.01$$

The adjustment factor K is given by:

$$K = \frac{400}{394.01} \approx 1.015$$

We will multiply each seasonal index by K:

$$\begin{aligned}
\text{Adjusted Index for Q1} &= 96.97 \times 1.015 \approx 98.66 \\
\text{Adjusted Index for Q2} &= 96.15 \times 1.015 \approx 97.62 \\
\text{Adjusted Index for Q3} &= 97.94 \times 1.015 \approx 99.27 \\
\text{Adjusted Index for Q4} &= 102.95 \times 1.015 \approx 104.21
\end{aligned}$$

Thus, the final adjusted seasonal indices are:

Quarter	**Adjusted Seasonal Index**
$Q1$	98.66
$Q2$	97.62
$Q3$	99.27
$Q4$	104.21

Conclusion

The Ratio to Moving Average method provides a systematic approach to estimating seasonal variations in time series data. In this example, we calculated the centered moving averages, expressed the original data as percentages, averaged these percentages to find the seasonal indices, and finally adjusted these indices to sum to a total of 400. This method enables businesses to better understand seasonal effects and make informed decisions based on these insights.

Example: 2

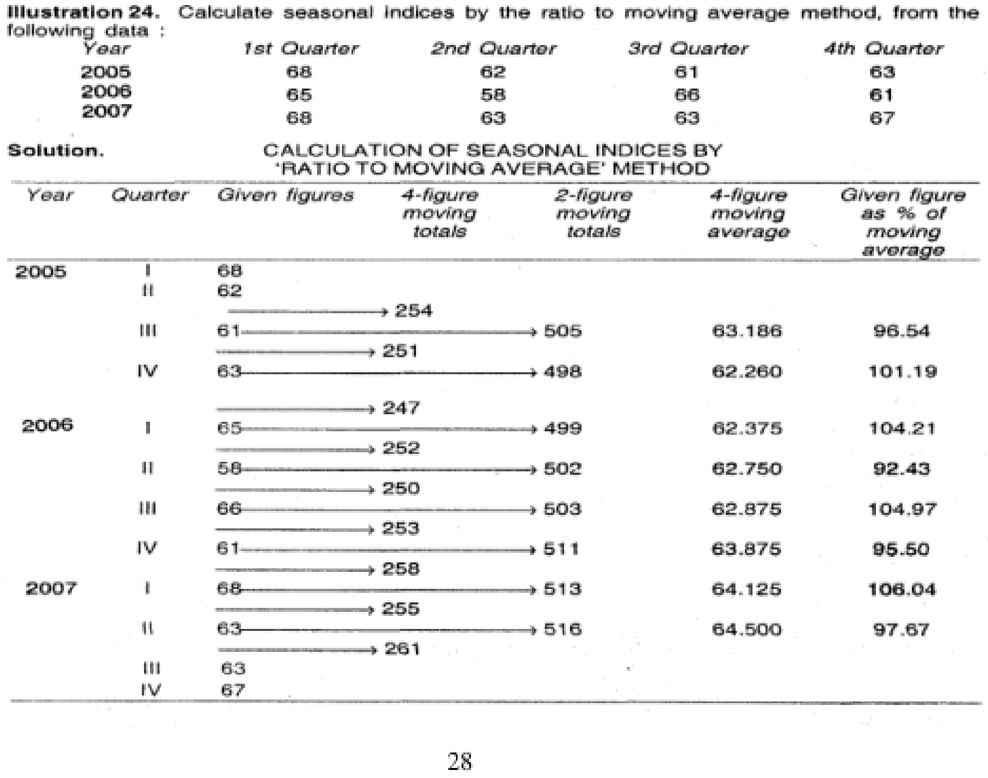


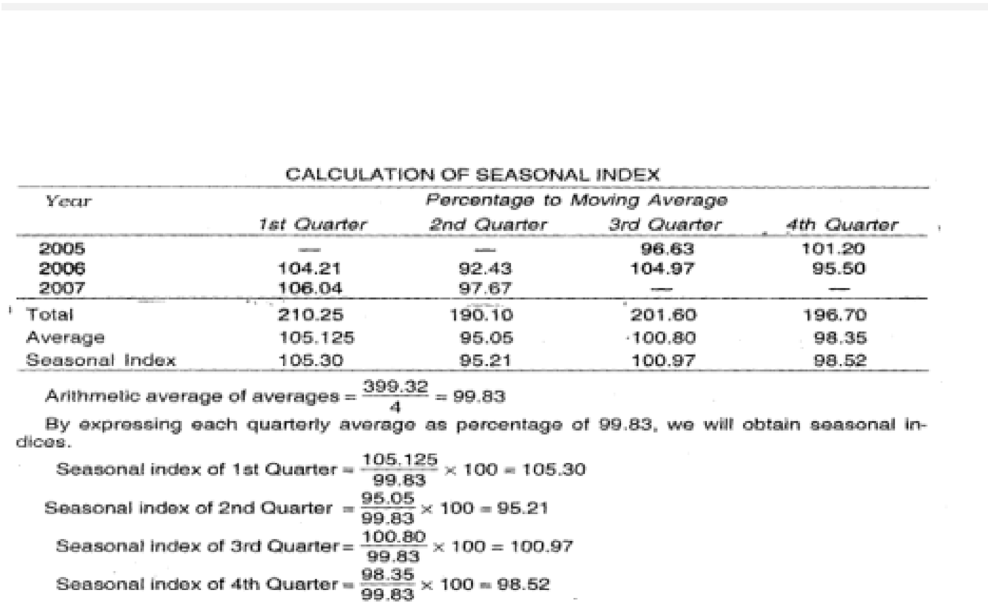
Illustration 24. Calculate seasonal indices by the ratio to moving average method, from the following data :

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2005	68	62	61	63
2006	65	58	66	61
2007	68	63	63	67

Solution. CALCULATION OF SEASONAL INDICES BY 'RATIO TO MOVING AVERAGE' METHOD

Year	Quarter	Given figures	4-figure moving totals	2-figure moving totals	4-figure moving average	Given figure as % of moving average
2005	I	68				
	II	62				
			254			
	III	61		505	63.186	96.54
			251			
	IV	63		498	62.260	101.19
			247			
2006	I	65		499	62.375	104.21
			252			
	II	58		502	62.750	92.43
			250			
	III	66		503	62.875	104.97
			253			
	IV	61		511	63.875	95.50
			258			
2007	I	68		513	64.125	106.04
			255			
	II	63		516	64.500	97.67
			261			
	III	63				
	IV	67				

28



CALCULATION OF SEASONAL INDEX

Year	Percentage to Moving Average			
	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2005	—	—	96.63	101.20
2006	104.21	92.43	104.97	95.50
2007	106.04	97.67	—	—
Total	210.25	190.10	201.60	196.70
Average	105.125	95.05	·100.80	98.35
Seasonal Index	105.30	95.21	100.97	98.52

Arithmetic average of averages $= \frac{399.32}{4} = 99.83$

By expressing each quarterly average as percentage of 99.83, we will obtain seasonal indices.

Seasonal index of 1st Quarter $= \frac{105.125}{99.83} \times 100 = 105.30$

Seasonal index of 2nd Quarter $= \frac{95.05}{99.83} \times 100 = 95.21$

Seasonal index of 3rd Quarter $= \frac{100.80}{99.83} \times 100 = 100.97$

Seasonal index of 4th Quarter $= \frac{98.35}{99.83} \times 100 = 98.52$

5.4 Link relative method

The Link Relative Method, also known as Pearson's Method, is a systematic approach for measuring seasonal variations. The steps involved in this method are as follows:

1. Calculate the **Link Relatives** for each period using the formula:

$$\text{Link Relative for any period} = \frac{\text{Current period's figure}}{\text{Previous period's figure}} \times 100$$

2. Calculate the average of the Link Relatives for each period across all years using either the mean or median.

3. Convert the average Link Relatives into **Chain Relatives** based on the first season. The Chain Relative for any period is obtained as:

$$\text{Chain Relative for the first period} = 100$$

$$\text{Chain Relative for any period} = \frac{\text{Average Link Relative for that period} \times \text{Chain Relative of the previous period}}{100}$$

4. Compute the **Adjusted Chain Relatives** by subtracting the correction factor k_d from the $(k+1)$th Chain Relative, where $k = 1, 2, \ldots, 11$ for monthly data and $k = 1, 2, 3$ for quarterly data. The correction factor k_d is defined as:

$$k_d = \frac{100}{N}$$

where N denotes the number of periods (i.e., $N = 12$ for monthly data and $N = 4$ for quarterly data).

5. Finally, calculate the average of the corrected Chain Relatives and convert these values into percentages based on this average. These percentages represent the seasonal indices calculated by the Link Relative Method.

Advantages

- The Link Relative Method utilizes the data more effectively compared to the moving average method.

Disadvantages

- This method involves extensive calculations and is more complex than the moving average method.
- The average of Link Relatives may contain both trend and cyclical components, which are eliminated by applying corrections.

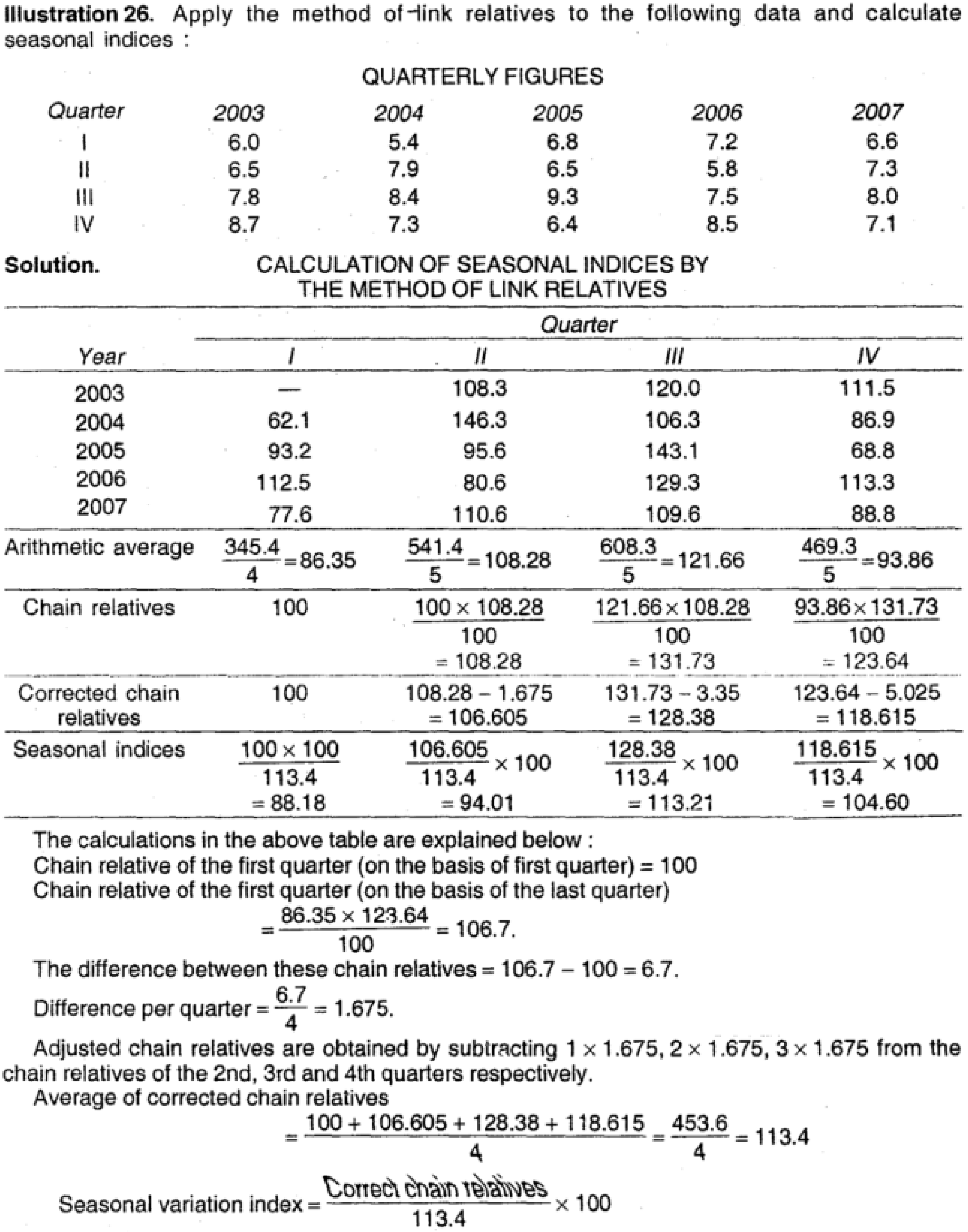


Illustration 26. Apply the method of link relatives to the following data and calculate seasonal indices :

QUARTERLY FIGURES

Quarter	*2003*	*2004*	*2005*	*2006*	*2007*
I	6.0	5.4	6.8	7.2	6.6
II	6.5	7.9	6.5	5.8	7.3
III	7.8	8.4	9.3	7.5	8.0
IV	8.7	7.3	6.4	8.5	7.1

Solution. CALCULATION OF SEASONAL INDICES BY THE METHOD OF LINK RELATIVES

	Quarter			
Year	*I*	*II*	*III*	*IV*
2003	—	108.3	120.0	111.5
2004	62.1	146.3	106.3	86.9
2005	93.2	95.6	143.1	68.8
2006	112.5	80.6	129.3	113.3
2007	77.6	110.6	109.6	88.8
Arithmetic average	$\frac{345.4}{4}=86.35$	$\frac{541.4}{5}=108.28$	$\frac{608.3}{5}=121.66$	$\frac{469.3}{5}=93.86$
Chain relatives	100	$\frac{100\times108.28}{100}$ $=108.28$	$\frac{121.66\times108.28}{100}$ $=131.73$	$\frac{93.86\times131.73}{100}$ $=123.64$
Corrected chain relatives	100	$108.28-1.675$ $=106.605$	$131.73-3.35$ $=128.38$	$123.64-5.025$ $=118.615$
Seasonal indices	$\frac{100\times100}{113.4}$ $=88.18$	$\frac{106.605}{113.4}\times100$ $=94.01$	$\frac{128.38}{113.4}\times100$ $=113.21$	$\frac{118.615}{113.4}\times100$ $=104.60$

The calculations in the above table are explained below :

Chain relative of the first quarter (on the basis of first quarter) = 100

Chain relative of the first quarter (on the basis of the last quarter)

$$=\frac{86.35\times123.64}{100}=106.7.$$

The difference between these chain relatives = 106.7 – 100 = 6.7.

Difference per quarter $=\frac{6.7}{4}=1.675.$

Adjusted chain relatives are obtained by subtracting 1 × 1.675, 2 × 1.675, 3 × 1.675 from the chain relatives of the 2nd, 3rd and 4th quarters respectively.

Average of corrected chain relatives

$$=\frac{100+106.605+128.38+118.615}{4}=\frac{453.6}{4}=113.4$$

$$\text{Seasonal variation index}=\frac{\text{Correct chain relatives}}{113.4}\times100$$

5.5 Cyclical and Random Fluctuations

Cyclical fluctuations refer to the oscillations in data that occur in a systematic pattern over a longer period, typically aligned with economic or business cycles. These cycles are not fixed in length and can vary widely, commonly seen in economic indicators like GDP, employment rates, and business profits.

For example, economic activity tends to rise during expansions and fall during recessions, creating a cyclical pattern. Cycles can span several years, such as the typical business cycle of about 4 to 10 years.

5.5.1 Example of Cyclical Fluctuations

Consider the following quarterly GDP data for a hypothetical economy over five years:

Year	$Q1$	$Q2$	$Q3$	$Q4$
1	200	220	230	240
2	250	270	260	280
3	290	300	310	320
4	310	330	340	350
5	340	350	360	370

In this example, we can observe an increasing trend in GDP, but there may also be fluctuations that correspond to economic cycles, indicating periods of growth followed by stagnation or decline.

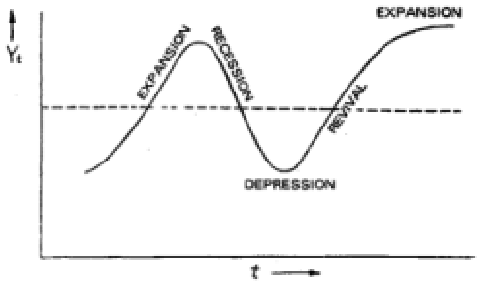


Figure 15: Cyclic Variations

article amsmath amssymb graphicx array

Methods for Measuring Cyclical Variations

Cyclical variations in data occur due to periodic fluctuations in economic activity, which can affect various sectors of the economy. Several methods are used to measure these variations effectively. The key methods include:

- Residual Method
- Reference Cycle Analysis Method
- Direct Method
- Harmonic Analysis Method

1. Residual Method

The Residual Method involves isolating the cyclical component of a time series by removing the trend and seasonal components. The cyclical variation is derived as the residuals after fitting a trend line and seasonal pattern to the data.

Steps:

1. Fit a trend line (linear, polynomial, etc.) to the time series data.
2. Identify and remove seasonal variations.
3. Calculate the residuals, which represent the cyclical variations.

Example:

Given a time series data of quarterly sales figures:

$$\text{Sales} = [200, 220, 210, 240, 260, 280, 270, 300]$$

Assume we fit a linear trend:

$$\text{Trend} = 180 + 10t \quad (t = 1, 2, \ldots, 8)$$

The fitted values and residuals can be calculated as follows:

Quarter	Sales	Trend	Residuals
1	200	190	10
2	220	200	20
3	210	210	0
4	240	220	20
5	260	230	30
6	280	240	40
7	270	250	20
8	300	260	40

The residuals represent the cyclical variations.

2. Reference Cycle Analysis Method

The Reference Cycle Analysis Method involves comparing a specific cycle with a standard reference cycle. The cyclical components of different time series can be analyzed to identify similarities and differences against a benchmark.

Steps:

1. Define a reference cycle based on historical data.
2. Compare the current data cycle with the reference cycle.
3. Measure deviations and similarities quantitatively.

Example:

Suppose the reference cycle is defined as follows:

$$\text{Reference Cycle} = [1, 0.9, 1.1, 1.2]$$

We compare this with a new cycle:

$$\text{Current Cycle} = [1.1, 0.8, 1.2, 1.3]$$

The deviations can be calculated:

$$\text{Deviation} = \frac{\text{Current Cycle}}{\text{Reference Cycle}} = [1.1, 0.89, 1.09, 1.08]$$

This analysis helps in assessing the performance against established benchmarks.

3. Direct Method

The Direct Method involves directly measuring the cyclical component from the time series data without removing the trend or seasonal effects. This method focuses on identifying peaks and troughs in the data.

Steps:

1. Identify the peaks and troughs in the data.
2. Calculate the amplitude of the cycles (the difference between peaks and troughs).
3. Analyze the duration of cycles to assess periodicity.

Example:

Given a time series of monthly sales:

$$\text{Sales} = [100, 120, 130, 125, 150, 160, 140, 130]$$

Identifying peaks and troughs:

$$\text{Peaks} = [130, 150, 160] \quad \text{Troughs} = [100, 125, 140]$$

The amplitudes can be calculated as follows:

$$\text{Amplitude} = \text{Peak} - \text{Trough}$$

For the first cycle:

$$\text{Amplitude} = 130 - 100 = 30$$

And so forth for each cycle.

4. Harmonic Analysis Method

Harmonic Analysis Method is a mathematical technique used to decompose time series data into its constituent sine and cosine components. This method is effective in identifying cyclical patterns in data that may not be readily apparent.

Steps:

1. Use Fourier transforms to convert the time series data into the frequency domain.
2. Identify significant harmonics that represent cyclical variations.
3. Reconstruct the cyclical component using selected harmonics.

Example:

Suppose we have a time series data:

$$\text{Data} = [1, 2, 3, 4, 5, 4, 3, 2]$$

Applying Fourier transform yields harmonics:

$$\text{Harmonics} = [A_1 \sin(\omega_1 t), A_2 \cos(\omega_2 t), \ldots]$$

Assuming two significant harmonics:

$$\text{Cyclical Component} = 2 \sin\left(\frac{2\pi}{T} t\right) + 3 \cos\left(\frac{2\pi}{T} t\right)$$

Where T is the period of the cycle.

The reconstructed cycle can show periodic behavior that highlights cyclical variations.

Conclusion

Each of these methods provides unique insights into the cyclical variations in time series data. The choice of method depends on the nature of the data, the underlying cycles, and the objectives of the analysis.

5.6 Random Fluctuations

Random fluctuations are unpredictable variations in data that do not follow a discernible pattern. These fluctuations can be caused by irregular, unforeseen events such as natural disasters, political instability, or sudden market changes. Unlike cyclical fluctuations, random variations are typically short-term and do not contribute to the overall trend.

5.6.1 Example of Random Fluctuations

Consider a stock price over time that displays random fluctuations due to market sentiment, news events, or economic reports:

Day	Stock Price
1	100
2	98
3	102
4	101
5	95
6	110
7	97
8	105
9	99
10	103

In this example, the stock price fluctuates randomly without showing any consistent trend, indicating that the changes are primarily driven by unpredictable market forces.

5.6.2 Deseasonalisation

Deseasonalisation is the process of removing seasonal components from time series data to obtain data that reflects only the underlying trends and cycles. The resulting data, free from seasonal variations, is known as **deseasonalised data**.

1. Multiplicative Model

In a multiplicative model, the relationship between the observed data Y_t, the trend T_t, and the seasonal component S_t is given by:

$$Y_t = T_t \times S_t$$

To deseasonalise the data, we divide the original data by the seasonal index. The seasonal index is typically expressed as a percentage, so we must adjust for that by using an adjustment multiplier of 100.

Formula for Deseasonalisation:

$$\text{Deseasonalised Data} = \frac{Y_t}{\text{Seasonal Index} \times 0.01}$$

Example:

Consider the following quarterly sales data and corresponding seasonal indices:

Quarter	Sales (Y)	Seasonal Index
$Q1$	120	110
$Q2$	150	90
$Q3$	180	100
$Q4$	200	130

Calculating the deseasonalised data:

$$\text{Deseasonalised Sales (Q1)} = \frac{120}{110 \times 0.01} = \frac{120}{1.1} \approx 109.09$$

$$\text{Deseasonalised Sales (Q2)} = \frac{150}{90 \times 0.01} = \frac{150}{0.9} \approx 166.67$$

$$\text{Deseasonalised Sales (Q3)} = \frac{180}{100 \times 0.01} = \frac{180}{1.0} = 180.00$$

$$\text{Deseasonalised Sales (Q4)} = \frac{200}{130 \times 0.01} = \frac{200}{1.3} \approx 153.85$$

The deseasonalised data is:

Quarter	Deseasonalised Sales
$Q1$	109.09
$Q2$	166.67
$Q3$	180.00
$Q4$	153.85

2. Additive Model

In an additive model, the relationship is expressed as:

$$Y_t = T_t + S_t$$

In this case, deseasonalisation involves subtracting the seasonal component from the original data.

Formula for Deseasonalisation:

$$\text{Deseasonalised Data} = Y_t - S_t$$

Example:

Using the same quarterly sales data:

Quarter	Sales (Y)	Seasonal Component (S)
$Q1$	120	10
$Q2$	150	20
$Q3$	180	30
$Q4$	200	40

Calculating the deseasonalised data:

$$\text{Deseasonalised Sales (Q1)} = 120 - 10 = 110$$

$$\text{Deseasonalised Sales (Q2)} = 150 - 20 = 130$$

$$\text{Deseasonalised Sales (Q3)} = 180 - 30 = 150$$

$$\text{Deseasonalised Sales (Q4)} = 200 - 40 = 160$$

The deseasonalised data is:

Quarter	Deseasonalised Sales
$Q1$	110
$Q2$	130
$Q3$	150
$Q4$	160

Uses and Limitations of Seasonal Indices

Uses:

- Seasonal indices provide a quantitative measure of typical seasonal behavior.
- They are used for forecasting and making informed business decisions by understanding seasonal fluctuations.

Limitations:

- Seasonal indices may not capture unexpected shocks or anomalies in data.
- They rely on historical data, which may not always predict future patterns accurately.

Conclusion

Deseasonalisation is a critical step in time series analysis, allowing analysts to focus on the underlying trends and cycles in data without the influence of seasonal fluctuations.

5.7 Variate Difference Methods

Variate difference methods, also known as difference methods or differencing, are techniques used in time series analysis to stabilize the mean of a time series by removing changes in the level of a time series, thereby making it stationary. This is particularly useful for analyzing seasonal data or trends.

5.7.1 Example 1: Monthly Sales Data

Consider a small business that records its monthly sales over six months as follows:

Month	Sales (in thousands)
1	50
2	60
3	70
4	80
5	65
6	75

To apply the variate difference method, we will calculate the first differences of the sales data.

Step 1: Calculate First Differences

The first difference is calculated as:

$$D_t = Y_t - Y_{t-1}$$

Where D_t is the first difference at time t and Y_t is the sales at time t.

Month	Sales (Y)	First Difference (D)
1	50	–
2	60	$60 - 50 = 10$
3	70	$70 - 60 = 10$
4	80	$80 - 70 = 10$
5	65	$65 - 80 = -15$
6	75	$75 - 65 = 10$

Step 2: Analyze the Differences

The first differences show that the sales increased consistently for the first four months but decreased in the fifth month. The last month, however, saw an increase again. This information helps the business understand that although sales fluctuated, the overall trend was increasing with occasional drops.

5.7.2 Example 2: Daily Temperature Records

Suppose we have daily temperature records for a week as follows:

Day	Temperature (°C)
1	20
2	22
3	24
4	23
5	25
6	27
7	26

We will apply the variate difference method to analyze the temperature changes.

Step 1: Calculate First Differences

Using the same formula for first differences:

$$D_t = Y_t - Y_{t-1}$$

We calculate the first differences:

Day	Temperature (Y)	First Difference (D)
1	20	–
2	22	$22 - 20 = 2$
3	24	$24 - 22 = 2$
4	23	$23 - 24 = -1$
5	25	$25 - 23 = 2$
6	27	$27 - 25 = 2$
7	26	$26 - 27 = -1$

Step 2: Analyze the Differences

From the first differences, we see that the temperature generally increased over the week, with minor fluctuations on days 4 and 7. The differences provide insight into the daily temperature variations, indicating stability with minor drops.

Differencing

The simplest form of variate difference methods is first-order differencing, where the difference between consecutive observations is calculated. The first difference is given by:

$$\Delta Y_t = Y_t - Y_{t-1}$$

Where: - Y_t is the value at time t, - Y_{t-1} is the value at the previous time period.

Example of First-Order Differencing

Consider the following time series data representing monthly sales figures (in thousands):

Month	Sales
1	100
2	120
3	130
4	150
5	180

The first-order differences can be calculated as follows:

Month	Sales	ΔY_t
1	100	–
2	120	$120 - 100 = 20$
3	130	$130 - 120 = 10$
4	150	$150 - 130 = 20$
5	180	$180 - 150 = 30$

The resulting first-order differences:

Month	ΔY_t
2	20
3	10
4	20
5	30

Second-Order Differencing

If the time series still shows non-stationarity after first differencing, a second-order differencing can be applied:

$$\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1}$$

This method is useful in capturing the cyclical patterns that may remain even after the first differencing.

Example of Second-Order Differencing

Continuing from the previous example, we calculate the second-order differences:

Month	ΔY_t	$\Delta^2 Y_t$
2	20	–
3	10	$10 - 20 = -10$
4	20	$20 - 10 = 10$
5	30	$30 - 20 = 10$

The second-order differences indicate the rate of change of the first differences, helping us understand the underlying dynamics of the time series data.

5.7.3 Conclusion

The variate difference methods are effective in analyzing trends and fluctuations in time series data. By calculating and interpreting the differences, we can derive valuable insights into the underlying patterns in the data.

Question Bank

1. Distinguish between seasonal variations and cyclical fluctuations. How would you measure secular trend in any given data?
2. Describe the method of link relatives for calculating the seasonal variation indices.
3. How would you determine seasonal variation in the absence of trend?
4. Briefly describe the relative merits and demerits of the ratio to trend and ratio to moving average methods.
5. What do you understand by cyclical fluctuations in time series?
6. What do you understand by random fluctuation in time series?
7. Explain the term "Business cycle" and point out the necessity of its study in time series analysis.
8. Calculate seasonal variation for the following data of sales in thousands Rs. of a firm by the Ratio to trend method.

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1979	30	40	36	34
1980	34	52	50	44
1981	40	58	54	48
1982	52	76	68	62

9. Calculate seasonal indices by the Ratio to moving average method from the following data.

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1980	75	60	54	59
1981	86	65	63	80
1982	90	72	66	85
1983	100	78	72	93

10. The data below gives the average quarterly prices of a commodity for five years. Calculate seasonal indices by the method of link relatives.

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1979	30	26	22	31
1980	35	28	22	36
1981	31	29	28	32
1982	31	31	25	35
1983	34	36	26	33

Module - 3

Chapter - 1

6 Index Numbers and their Definitions

6.1 Construction and Uses of Fixed and Chain based Index Numbers

6.2 Simple and Weighted Index Numbers

6.3 Laspeyres, Paasche's, Fisher's, and Marshall - Edgeworth Index Numbers

6.4 Optimum Tests for Index Numbers

6.5 Cost of Living Index Numbers

Module - 3

Chapter - 2

7 Forecasting Strategies

7.1 Leading variables and associated variables

7.2 Bass Model

7.3 Exponential Smoothing and Holt-Winters method

Module - 4

Chapter - 1

8 Basic Stochastic Models

8.1 White Noise, Random Walks, Fitted models & diagnostic plots

8.2 Autoregressive models

8.2.1 stationary and non-stationary Autoregressive process

Module - 4

Chapter - 2

9 Time series Regression and Exploratory Data Analysis

9.1 Classical Regression

9.2 Exploratory Data Analysis

9.3 generalized least square method

9.4 linear models with seasonal variables

9.5 Harmonic seasonal models

9.6 logarithmic transforms

Module - 5

Chapter - 1

10 Linear Models

10.1 Moving Average models

10.2 Fitted MA Models

10.2.1 Autoregressive Moving Average Models

10.3 Differential Equations

10.4 Autocorrelation and Partial Correlation

10.5 Forecasting & Estimation

10.6 Non-stationary Models

10.6.1 Building non-seasonal ARIMA Models

10.6.2 ARCH Models & GARCH Models